CHIMERIC AUTOPROCESSING POLYPEPTIDES AND USES THEREOF

FIELD AND BACKGROUND OF THE INVENTION

The present invention relates to polypeptides having the capacity to display auto-cleavage, polynucleotides encoding such polypeptides, and uses of such polypeptides and polynucleotides for reversibly binding proteins to specific substrates, reversibly binding specific substrates to each other, and for splicing amino acid sequences. More particularly, the present invention relates to chimeric polypeptides capable of auto-cleaving at defined locations, including auto-cleaving resulting in defined auto-splicing, to polynucleotides suitable for expressing such polypeptides, and to methods of using such polypeptides and polynucleotides for protein purification, affinity selection of display phages, and post-translational ligation of proteins.

Autoprocessing protein domains, such as inteins and Hogs, have the capacity to post-translationally auto-cleave or auto-splice flanking polypeptide sequences and thereby serve as unique and potent protein engineering tools useful in various applications, including protein purification, affinity selection of display phages, generation of cytotoxic proteins, segmental modification or labeling of proteins, protein or peptide cyclization, and generation of reactive polypeptide termini in expressed proteins for various biochemical reactions, including protein ligation (Perler and Adam, 2000. Curr Opin Biotechnol. 11, 377-83). However, the usefulness of the presently available repertoire of autoprocessing polypeptides is hampered by various limitations, as described in further detail hereinbelow.

Inteins are internal protein domains naturally occurring in a variety of host proteins (Hirata *et al.*, 1990. J. Biol. Chem. 265, 6726-6733; Kane *et al.*, 1990. Science 250, 651-657; Perler *et al.*, 1994. Nucl. Acids Res. 22, 1125-1127; Noren *et al.*, 2000. Angew. Chem. Int. Ed. 39, 450). Inteins have been found in organisms from all three domains of life, including in yeast and algal chloroplasts (eukaryotes), mycobacteria and cyanobacteria (bacteria), and thermophilic archaea (archaea). So far, no essential biological role has been shown for inteins, and all of their identified functions involve their own preservation and maintenance, with no apparent benefit to the host protein and organism (reviewed in Pietrokovski, 2000. Trends in Genetics 17,

465-472). At least some inteins are multifunctional, being able to both catalyze their own protein splicing and to home a copy of their gene into intein-less alleles (Gimble and Thorner, 1992. Nature 357, 301; Chong *et al.*, 1996. J Biol Chem. 271, 22159). Hogs are protein domains found in Hedgehogs which are proteins composed of an amino terminal Hedge protein domain and a carboxy terminal Hog protein region (Aspock G., 1999. Genome Res. 9, 909; Hammerschmidt *et al.*, 1997. Trends Genet. 13, 14). Other protein domains, such as various *Caenorhabditis elegans* carboxy terminal domains, are believed to autocatalytically cleave themselves from host proteins, thereby modulating the activity of the amino terminal parts (Burglin, 1996. Curr Biol. 6, 1047; Porter *et al.*, 1996. Cell 86, 21), similarly to Hogs.

Members of the intein and Hog protein domain families share the capacity to autocatalytically cleave the peptide bond joining them to polypeptides flanking their amino terminal ends ("amino terminal cleavage"). Inteins have the further capacity to cleave the peptide bond joining them to polypeptides flanking their carboxy terminal ends ("carboxy terminal cleavage") while splicing polypeptides flanking their amino and carboxy terminal ends (termed "exteins"), resulting in self-excision of the intein from the host protein, and concomitant ligation of the flanking extein domains with a peptide bond. Thus, intein-containing host proteins undergo a switch from an intein-containing state to an intein-less state via such a process. Most reported inteins furthermore also contain an endonuclease domain whose function is to mediate the copying of the intein gene into specific unoccupied genomic insertion points, thereby enabling intein propagation.

Both inteins and Hogs share a similar structure fold and contain characteristic "Hint" consensus motifs which mediate the biochemical reactions involved in the autocatalytic activities of these protein domains (Hall TM., 1997. Cell 91, 85; Pietrokovski S., 1994. Protein Sci. 3, 2340; Pietrokovski S., 1998. Protein Sci. 7, 64; Paulus, 2000. Annu. Rev. Biochem. 69, 447). These Hint motif-mediated biochemical reactions are similar in both inteins and Hogs, but are involved in different biological processes (Dalgaard *et al.*, 1997. J Comput Biol. 4, 193; Hall *et al.*, 1997. Cell 91, 85; Pietrokovski S., 1998. Protein Sci. 7, 64; Xu and Perler, 1996. EMBO J. 15, 5146). The initial biochemical reactions of intein and Hog amino terminal cleavage are identical; the peptide bond attaching the amino terminal end of the Hint domain to an

amino terminal-flanking sequence is converted into a thioester (or ester) bond, a trans-esterification reaction then covalently attaches the sequence flanking the carboxy terminal end of the intein, or a cholesterol molecule in the case of Hog proteins, to the amino terminal flanking sequence, thereby cleaving the bond attaching the amino terminal sequence to the Hint domain. In a process essential for organismal development, Hint-mediated autocatalytic excision of the carboxy terminal Hog protein domain from the amino terminal Hedge protein domain in Hedgehogs leads to covalent attachment of a cholesterol molecule to the carboxy end of the Hedge domain, leading to its activation and secretion from the cell (Porter JA. *et al.*, 1996. Cell 86, 21; Porter, JA. *et al.*, 1996. Science 274, 255). In the case of inteins, protein splicing is effected sequentially by cleavage of the bond attaching the intein amino terminal end to the carboxy terminal extein, ligation of the amino and carboxy terminal exteins, and cleavage of the bond attaching the intein carboxy terminal end to the carboxy terminal extein.

Mechanistic studies have determined the roles of highly conserved residues positioned near the intein/extein junctions in the splicing reaction (Chong *et al.*, 1996. J. Biol. Chem. 271, 22159-22168; Xu *et al.*, 1996. EMBO J. 15, 5146-5153; Stoddard *et al.*, 1998. Nat. Struct. Biol. 5, 3). These residues include: the Cys, Ser or Thr residue forming the amino terminal end of the intein, which initiates splicing with an acyl shift; the conserved Cys, Ser or Thr residue flanking the carboxy terminal end of the intein, which ligates the exteins through nucleophilic attack; and the conserved Asn forming the carboxy terminal end of the intein, which releases the intein from the ligated exteins via succinimide formation. The amino terminal acyl shift and the carboxy terminal succinimide formation cleavage activities of the intein are separable. The amino terminal cleavage takes place in two separate steps. In the first step, as described above, the peptide bond between the intein and the amino terminal extein is converted to a thioester (or ester in some cases). In the second step, the thioester bond is cleaved by a nucleophilic attack from the side-chain of the residue flanking the carboxy terminal end of the intein, causing a transesterification reaction.

Because the structural information required for splicing exists entirely within inteins, and since the process of splicing has no energy requirements (for example hydrolysis of ATP), such protein domains can be used in a variety of applications

involving intein insertion into foreign contexts. Various methods have been used in attempts to control and alter intein-mediated functions. Since endonuclease activity is not required for protein splicing, mini-inteins with accurate splicing activity have been generated by deletion of this central domain (Derbyshire *et al.*, 1997. Proc. Natl. Acad. Sci. USA. 94, 11466; Chong *et al.*, 1997. J. Biol. Chem. 272, 15587; and Shingledecker *et al.*, 1998. Gene 207, 187). Also, mutation of residues near the intein/extein junctions has been used to alter intein activity, for example, to yield isolated cleavage at one or both of the intein-extein junctions (Chong *et al.*, 1998. J. Biol. Chem. 273, 10567).

Thus, the ability to modulate the function of autoprocessing polypeptides such as inteins has broad potential application, as described above. In the case of protein purification where an autoprocessing polypeptide is used in conjunction with an affinity group to purify a desired target protein (Chong *et al.*, 1997. Gene 192, 271–281; Chong *et al.*, 1998. Nucl. Acids Res. 26, 5109), purification of a target protein is effected by co-expressing the target protein as a fusion protein containing a purification tag in one terminal segment, an internal autoprocessing polypeptide, and a target protein forming the other terminal segment. Such fusion proteins are exposed to affinity purification matrices designed to capture the tagged molecule. The target protein is then selectively released from the purification matrix by inducing autoprocessing polypeptide-mediated auto-cleavage of the peptide bond attaching the target protein to the autoprocessing polypeptide. Such a procedure is advantageous since autoprocessing polypeptide cleavage affects the fusion protein only, and thus non-specifically bound contaminant proteins are not released into the product stream. Furthermore, such a method does not employ contaminating and expensive proteases, such as those used in technologies employing protease-mediated cleavage of purification-tagged target proteins. The aforementioned strategy forms the basis of the protein purification systems such as the commercially available IMPACT-CN system (New England Biolabs, Beverly, MA).

However, prior art methods of using such autoprocessing polypeptides for applied uses have numerous drawbacks. In applied systems such as IMPACT-CN, the accessory molecule involved in cleavage of the thioester bond between the intein and the extein following amino terminal cleavage must be effected with a strong thiol-

containing nucleophile such as 2-mercaptoethanol or dithiothreitol (DTT), both of which are strong reducing agents which modify the carboxy terminal end of the extein. In such systems, although initial thioester formation is mediated by the intein, the actual cleavage of the extein is effected via non-enzymatic chemical cleavage of a thioester bond by a small nucleophilic molecule, thereby severely limiting the maximal reaction rates achievable. While such systems allow carboxy terminal cleavage, such cleavage has the drawback of resulting in undesirable amino terminal cleavage, thereby requiring the amino terminal fragment to be removed in an additional purification step. Furthermore, despite insights into intein structure and function, modifications often result in unacceptably low activity, poor precursor stability, or insolubility (Derbyshire *et al.*, 1997. Proc. Natl. Acad. Sci. USA. 94, 11466; Chong *et al.*, 1997. Gene 192, 271-281; Shingledecker *et al.*, 1998. Gene 207, 187; Chong *et al.*, 1998. Nucl. Acids Res. 26, 5109).

Thus, all prior art approaches have failed to provide an adequate solution for providing autoprocessing polypeptides optimal for protein engineering applications.

There is thus a widely recognized need for, and it would be highly advantageous to have, autoprocessing polypeptides devoid of the above limitation.

SUMMARY OF THE INVENTION

According to one aspect of the present invention there is provided an chimeric polypeptide comprising an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106, the polypeptide being capable of auto-cleavage.

According to another aspect of the present invention there is provided a polynucleotide encoding a chimeric polypeptide comprising an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,

60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106, the polypeptide being capable of auto-cleavage.

According to further features in preferred embodiments of the invention described below, the chimeric polypeptide further comprises an affinity tag capable of specifically binding a substrate.

According to still further features in preferred embodiments, the substrate is selected from the group consisting of a molecule, a compound, a virus, and a cell.

According to yet another aspect of the present invention there is provided a nucleic acid construct comprising a nucleic acid sequence encoding a chimeric polypeptide comprising an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106, the chimeric polypeptide being capable of auto-cleavage.

According to further features in preferred embodiments of the invention described below, the nucleic acid construct further comprises a promoter sequence being for directing expression of the chimeric polypeptide in an expression system.

According to still further features in preferred embodiments, the chimeric polypeptide further comprises an affinity tag capable of specifically binding a specific substrate.

According to still another aspect of the present invention there is provided a method of generating a chimeric polypeptide capable of displaying auto-cleavage, the method comprising generating a chimeric amino acid sequence including an autoprocessing segment, the autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106, thereby producing the

chimeric polypeptide capable of displaying auto-cleavage.

According to further features in preferred embodiments of the invention described below, the chimeric polypeptide includes an affinity tag.

According to a further aspect of the present invention there is provided a method of purifying a protein, the method comprising: (a) generating a chimeric polypeptide including an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106, the autoprocessing segment being terminally attached to, or flanked by, an amino acid sequence of the protein, the chimeric polypeptide being capable of auto-cleavage when subjected to suitable conditions to thereby remove the amino acid sequence of the protein from the chimeric polypeptide thereby generating the protein; (b) immobilizing the chimeric polypeptide to a support; and (c) subjecting the chimeric polypeptide to the suitable conditions, thereby purifying the protein.

According to further features in preferred embodiments of the invention described below, the method of purifying a protein further comprises the step of separating the protein from the autoprocessing segment following step (c).

According to still further features in preferred embodiments, the support includes an antibody or antibody fragment capable of specifically binding the autoprocessing segment, and the immobilizing is via the autoprocessing segment.

According to still further features in preferred embodiments, the chimeric polypeptide further includes an affinity tag sequence, and the immobilizing is via the affinity tag sequence.

According to still further features in preferred embodiments, the support includes a specific ligand of the affinity tag sequence, and the immobilizing is via the specific ligand of the affinity tag sequence.

According to still further features in preferred embodiments, the generating the chimeric polypeptide is effected by synthesizing a polynucleotide encoding the chimeric polypeptide and expressing the polynucleotide in an expression system.

According to still further features in preferred embodiments, the expression system is a cellular expression system or a cell-free expression system.

According to still further features in preferred embodiments, the cellular expression system is an *E. coli* cellular expression system.

According to still further features in preferred embodiments, the cell-free expression system is an *E. coli* S30 extract expression system.

According to still further features in preferred embodiments, the polynucleotide comprises a promoter sequence being for directing the expression of the chimeric polypeptide.

According to still further features in preferred embodiments, the promoter sequence is inducible by isopropyl beta-D-thiogalactoside.

According to still further features in preferred embodiments, the auto-cleavage results in auto-splicing.

According to still further features in preferred embodiments, the auto-splicing is auto-splicing of segments of the chimeric polypeptide flanking the autoprocessing segment.

According to a further aspect of the present invention there is provided a method of reversibly attaching a first substrate to a second substrate, the method comprising: (a) providing a chimeric polypeptide including an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106 flanked by a first amino acid sequence capable of binding the first substrate and a second amino acid sequence capable of binding the second substrate, the chimeric polypeptide being capable of auto-cleavage when subjected to suitable conditions, to thereby release the first amino acid sequence from the second amino acid sequence; (b) exposing the first substrate and the second substrate to the chimeric polypeptide, thereby generating a complex including the first substrate attached via the chimeric polypeptide to the second substrate; and (c) subjecting the complex to the suitable conditions, thereby detaching the first substrate from the second substrate.

According to further features in preferred embodiments of the invention described below, each of the first and second substrates is independently selected from the group consisting of a molecule, a compound, a virus, and a cell.

According to still further features in preferred embodiments, the molecule is amylose or chitin.

According to still further features in preferred embodiments, the virus is a bacteriophage.

According to still further features in preferred embodiments, the chimeric polypeptide includes an affinity tag sequence, and the binding the first substrate or the binding the second substrate is via the affinity tag sequence.

According to still further features in preferred embodiments, the affinity tag sequence is a maltose-binding domain or a chitin-binding domain.

According to still further features in preferred embodiments, the autoprocessing segment is selected from the group consisting of BIL1_cloth, BIL2_cloth, BIL3_cloth, BIL4_cloth, BIL5_cloth, BIL6_cloth, BIL7_cloth, BIL9_cloth, BIL10_cloth, BIL11_cloth, 3875_87_magma, FhaB_manha, BIL2_neigo, BIL3_neigo, BIL5_neigo, BIL6_neigo, MafB1_neigo, MafB2_neigo, B0369+_neimeB, B0372+_neimeB, B0655+_neimeB, A2115_neime, BIL2_neimeC, BIL3_neimeC, BIL4_neimeC, BIL5_neimeC, BIL6_neimeC, MafB1_neimeC, FhaB1_psefl-PfO-1, FhaB1_psefl-SBW25, FhaB_psesy, SCP1.201_strco, 39_9_thefus, BIL1_gemob, BIL2_gemob, 0709_lepin, 3725_lepin, 3719_lepin, o665_myxxa, o1078_myxxa, o1070_myxxa, BIL1_strav, BIL2_strav, BIL3_strav, BIL1_pirsp, BIL1_chrvi, BIL1_glovi, BIL2_glovi, BIL3_glovi, BIL4_glovi, BIL5_glovi, BIL6_glovi, BIL7_glovi, o649_versp, o5687_versp, o3395_versp, II0519_brume, BIL2_magma, BIL3_magma, BIL4_magma, BIL5_magma, BIL6_magma, 06786_metex, 00126_rhoca, 00199_rhoca, 00459_rhoca, 00460_rhoca, 00746_rhoca, 00949_rhoca, 01216_rhoca, 01374_rhoca, 01523_rhoca, 01524_rhoca, 02710_rhoca, 03530_rhoca, 4825_rhosp, BIL2_rhosp, BIL1_silpo, BIL2_silpo, BIL3_silpo, BIL4_silpo, BIL5_silpo, BIL6_silpo, BIL7_silpo, BIL8_silpo, BIL9_silpo, BIL10_silpo, BIL11_silpo, BIL12_silpo, BIL13_silpo, BIL14_silpo, BIL15_silpo, BIL16_silpo, Bil1_rhile, BIL1_unknwn, and BIL2_unknwn.

According to still further features in preferred embodiments, the

autoprocessing segment is derived from a protein of an organism belonging to a genus selected from the group consisting of *Brucella, Clostridium, Magnetospirillum, Mannheimia, Methylobacterium, Neisseria, Pseudomonas, Rhodobacter, Silicibacter, Streptomyces, Thermobifida, Rhizobium, Chromobacterium, Myxococcus, Leptospira, Pirellula, Gemmata, Gloeobacter* and *Verrucomicrobium*.

According to still further features in preferred embodiments, the organism is selected from the group consisting of *Rhodobacter capsulatus, Rhodobacter sphaeroides, Silicibacter pomeroyi, Brucella melitensis, Brucella suis, Magnetospirillum magnetotacticum, Methylobacterium extorquens, Rhizobium leguminosarum, Neisseria meningitidis, Neisseria meningitidis, Neisseria meningitidis, Neisseria gonorrhoeae, Chromobacterium violaceum, Pseudomonas syringae, Pseudomonas fluorescens, Pseudomonas fluorescens, Mannheimia haemolytica, Myxococcus xanthus, Leptospira interrogans, Streptomyces coelicolor, Streptomyces avermitilis, Thermobifida fusca, Clostridium thermocellum, Pirellula species 1, Gemmata obscuriglobus, Gloeobacter violaceus,* and *Verrucomicrobium spinosum*.

According to still further features in preferred embodiments, the auto-cleavage results in removal of a segment of the chimeric polypeptide adjacent to an amino terminal end or a carboxy terminal end of the autoprocessing segment.

According to still further features in preferred embodiments, the segment of the chimeric polypeptide adjacent to the autoprocessing segment is an amino terminal segment or a carboxy terminal segment of the chimeric polypeptide.

According to still further features in preferred embodiments, the segment of the chimeric polypeptide adjacent to the carboxy terminal end of the autoprocessing segment includes an amino acid residue comprising a nucleophilic group at an amino terminal end thereof.

According to still further features in preferred embodiments, the nucleophilic group is a hydroxyl group.

According to still further features in preferred embodiments, the amino acid residue is a threonine residue.

According to still further features in preferred embodiments, the segment of the chimeric polypeptide adjacent to the amino terminal end of the autoprocessing

segment includes a serine amino acid residue at a carboxy terminal end thereof.

According to still further features in preferred embodiments, the chimeric polypeptide is capable of the auto-cleavage under a condition selected from the group consisting of a temperature selected from a range of 33 °C to 41 °C, a pH selected from a range of pH 7.8 to pH 8.2, and a concentration of dithiothreitol selected from a range of 0.1 mM to 20 mM.

The present invention successfully addresses the shortcomings of the presently known configurations by providing novel chimeric polypeptides capable of defined auto-cleaving, including auto-cleaving resulting in defined auto-splicing, polynucleotides suitable for expressing such polypeptides, and methods of using such polypeptides and polynucleotides to purify proteins, affinity-select display phages, and post-translationally ligate proteins together.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the patent specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is herein described, by way of example only, with reference to the accompanying drawings. With specific reference now to the drawings in detail, it is stressed that the particulars shown are by way of example and for purposes of illustrative discussion of the preferred embodiments of the present invention only, and are presented in the cause of providing what is believed to be the most useful and readily understood description of the principles and conceptual aspects of the invention. In this regard, no attempt is made to show structural details of the invention in more detail than is necessary for a fundamental understanding of the invention, the description taken with the drawings making apparent to those skilled in

the art how the several forms of the invention may be embodied in practice.

In the drawings:

FIG. 1a is a sequence alignment diagram of the amino acid sequences of various Type A BIL domains (SEQ ID NOs: 8-62). The names of the domains and the amino acid sequence coordinates of their amino terminal residues in their host proteins are indicated to the left of the sequence. Numbers indicated in parentheses indicate the amino acid residue length of an intervening amino acid sequence which may have any amino acid sequence. Dashed lines have been inserted to assist in visualizing the alignments.

FIG. 1b is a sequence alignment diagram of the amino acid sequences of various Type B BIL domains (SEQ ID NOs: 63-104). The names of the domains and the amino acid sequence coordinates of their amino terminal residues in their host proteins are indicated to the left of the sequence. Numbers indicated in parentheses indicate the amino acid residue length of an intervening amino acid sequence which may have any amino acid sequence. Dashed lines have been inserted to assist in visualizing the alignments.

FIG. 2a is a sequence diagram depicting an amino acid sequence motif (SEQ ID NO: 105) exclusively defining amino acid sequences of a subset of Type A BIL domains, including: 39_9_thefus, SCP1.201_strco, 3875_87_magma, B0372+_neimeB, B0655+_neimeB, A2115_neime, MafB1_neimeC, BIL2_neimeC, BIL4_neimeC, BIL5_neimeC, BIL6_neimeC, MafB1_neigo, BIL2_neigo, BIL3_neigo, MafB2_neigo, BIL5_neigo, BIL6_neigo, FhaB_psesy, FhaB_manha, FhaB1_psefl-PfO-1, FhaB1_psefl-SBW25, BIL6_cloth, BIL5_cloth, BIL2_cloth, BIL4_cloth, BIL1_cloth, BIL8_cloth, BIL9_cloth, BIL1_gemob, BIL2_gemob, 0709_lepin, 3725_lepin, o1078_myxxa, o1070_myxxa, BIL1_strav, BIL2_strav, BIL3_strav, BIL1_pirsp, BIL1_chrvi, o3395_versp, o5687_versp, o649_versp, BIL1_glovi, BIL2_glovi, BIL3_glovi, and BIL4_glovi. Amino acid residues are indicated in standard single-letter code. X - any amino acid; X(1-100) - any amino acid sequence composed of 1 to 100 amino acid residues.

FIG. 2b is a sequence diagram depicting an amino acid sequence motif (SEQ ID NO: 106) exclusively defining amino acid sequences of a subset of Type B BIL domains, including: 4825_rhosp, BIL2_rhosp, 00588_rhoca, 02710_rhoca,

01524_rhoca, 01523_rhoca, 00126_rhoca, 01216_rhoca, 00949_rhoca, 01374_rhoca, 00459_rhoca, 00460_rhoca, 00746_rhoca, 03530_rhoca, 00199_rhoca, BIL3_magma, BIL4_magma, BIL1_brusu, BIL1_unknwn, BIL2_unknwn, 06786_metex, BIL1_silpo, BIL2_silpo, BIL3_silpo, BIL4_silpo, BIL5_silpo, BIL6_silpo, BIL7_silpo, BIL8_silpo, BIL9_silpo, BIL10_silpo, BIL11_silpo, BIL12_silpo, BIL13_silpo, BIL14_silpo, BIL15_silpo, BIL16_silpo, BIL1_rhile, and II0519_brume. Amino acid residues are indicated in standard single-letter code. X - any amino acid; X(1-100) - any amino acid sequence composed of 1 to 100 amino acid residues.

FIGs. 3a-z are block-logo diagrams depicting conserved amino acids of Hint-like motifs of Type A BIL domains (Figures 3a-g) and Type B BIL domains (Figures 3h-o) and of homologous Hint motifs of Hog proteins (Figures 3p-t) and inteins (Figures 3u-z). Unique BIL domain motifs are underlined with hatched lines. Motifs are ordered left to right in the amino to carboxy terminal positions along the protein sequences. Similar motifs are vertically aligned. The motifs are shown as sequence logos in which the heights of amino acid letter designations are proportional to their degree of conservation in each position. Protein splicing active site residues of intein Hint domains are indicated by asterisks. Motifs were identified and are displayed as previously described (Pietrokovski S., 1998. Protein Sci. 7, 64). The BIL domain motifs shown include sequences described in Figures 1a-b and Tables 1-2. The intein and hedgehog Hint domain sequences were obtained from previously published sources (Aspock G., 1999. Genome Res. 9, 909; Pietrokovski S., 2001. Trends Genet. 17, 465). The amino acid residue position shown forming the carboxy terminal end of the motif depicted in Figure 3z actually represents the amino acid residue forming the amino terminal end of the carboxy terminal extein. Only intein and hedgehog motifs common to Hint domains are shown.

FIGs. 4a-b are dendrograms depicting phylogenetic relationships of Type A and Type B BIL domains, respectively.

FIGs. 5a-b are autoradiographs depicting SDS-PAGE separation of *in-vitro* translated, [^{35}S]-methionine–labeled, protein products of the MBP-BIL-CBD expression constructs pC2C-PsyBIL, and pC2C-RspBIL2 (Figures 5a and 5b, respectively, "MBC" lane). Translation of control constructs pC2C ("MC" lane) was

used as a positive control and no DNA template ("–" lane) were used as a negative control. Molecular weights were estimated using translation products of pBEST*luc* as reference standards. In Figure 5a, molecular weights ("mw" lane) were further estimated using unlabeled protein markers (dotted lines). In Figure 5b [MBP-CBD]′ and [MBP]′ indicate splicing and amino terminal cleavage products derived from MBP-RspBIL2-CBD. These fragments migrated at molecular weights greater than those the corresponding control fragments due to their containing residual BIL2_rhosp-flanking amino acid residues as a result of the pC2C-RspBIL2 cloning scheme described under Materials and Methods. Expected molecular weights of identified fragments are indicated in parentheses. The expected molecular weights of MBP-PsyBIL-CBD and of its carboxy terminal cleavage product MBP-PsyBIL, and its MBP-CBD splicing product were 66.3 and 59.1, and 50.6 kDa, respectively.

FIG. 5c is a diagram of the amino acid sequence of FhaB_psesy (SEQ ID NO: 107) depicting the putative protein splicing motifs (underlined) and catalytic residues (double-underlined) responsible for auto-cleavage/-splicing activity by this Type A BIL domain.

FIGs. 6a-b are photographs depicting SDS-PAGE analysis of autocatalytically processed protein products overexpressed *in-vivo* in *E. coli* transformed with the MBP-PsyBIL-CBD ("MBC" lane) expression construct pC2C-PsyBIL. Figure 6a depicts Coomassie Blue staining of chitin (lane 2) or amylose (lane 3) affinity column separated protein. Lane 4 is a control showing protein from *E. coli* transformed with pC2C to overexpress MBP-CBD ("MC" lane) chimeric protein. Fragments corresponding to the MBP-PsyBIL carboxy terminal auto-cleavage product and the MBP-CBD auto-splicing product of the chimera are indicated. Figure 6b is an autoradiograph depicting Western immunoblotting analysis of separated protein following purification on chitin beads. Both chitin purified samples from *E. coli* transformed with pC2C-PsyBIL or control plasmid pC2C to overexpress MBP-PsyBIL-CBD or MBP-CBD chimeric proteins, respectively, were separated in duplicate lanes and blotted onto a single nitrocellulose membrane. The membrane was cut in half and each sample was reacted in duplicate with either anti-MBP (anti-M) or anti-CBD (anti-C) antibodies. Both anti-M and anti-C antibodies reacted with the protein band corresponding to the mass of the MBP-CBD product. Protein bands

corresponding to MBP-PsyBIL and MBP products, that appear following purification on chitin beads result from non specific binding by excess amounts of overexpressed protein. Expected molecular weights of identified fragments are indicated in parentheses. The expected molecular weights of MBP-PsyBIL-CBD and of its carboxy terminal cleavage product MBP-PsyBIL, and its MBP-CBD splicing product were 66.3 and 59.1, and 50.6 kDa, respectively.

FIGs. 6c-d are data plots depicting MALDI mass spectra of MBP-CBD ligation product (Figure 6c) and MBP-PsyBIL carboxy terminal cleavage product (Figure 6d) electroeluted from SDS-PAGE gels. The expected molecular weights of MBP-PsyBIL-CBD and of its carboxy terminal cleavage product MBP-PsyBIL, and its MBP-CBD splicing product were 66.3 and 59.1, and 50.6 kDa, respectively.

FIG. 7 is an amino acid sequence diagram depicting positioning of peptide sequences identified by MALDI peptide mass mapping analysis within the amino acid sequence of the MBP-CBD splicing product (SEQ ID NO: 108). Twenty-five tryptic peptide masses (underlined) were assigned to the amino acid sequence of the MBP-CBD protein, corresponding to 49 % coverage of the MBP-CBD sequence. Lettering in non-bold/italic font indicates the amino acid sequence of the MBP tag, and the amino acid sequence of the CBD tag (amino acids 394–461) is indicated in bold+italic font. The peptide corresponding to amino acids 388–396 contains the BIL domain splice site between amino acids Ser393 and Thr394.

FIG. 8 is an amino acid sequence diagram depicting MALDI peptide mapping of the 59.3 kDa MBP-PsyBIL carboxy terminal cleavage product (SEQ ID NO: 109). Underlined sequences correspond to peptides detected by MALDI. Lettering in non-bold/italic font indicates the amino acid sequence of the MBP tag and that in bold/italic font indicates the amino acid sequence of the PsyBIL domain. The carboxy terminal end of the protein, asparagine N541 represents the carboxy terminal of the PsyBIL domain. The expected molecular weights of MBP-PsyBIL-CBD and of its carboxy terminal cleavage product MBP-PsyBIL, and its MBP-CBD splicing product were 66.3 and 59.1, and 50.6 kDa, respectively.

FIG. 9 is a schematic diagram depicting functions of BIL domains. Hint domains are shown as dark gray horseshoes with their flanks as ovals. Proteins are depicted with amino termini positioned on the left.

FIGs. 10a-c are electrophoretic analyses depicting C-terminal auto-cleavage by MBP-RspBIL2a-CBD chimera. Figure 10a is a photograph of a Coomassie blue stained electrophoretic separation of *in-vivo* expressed MBP-RspBIL2a-CBD chimera affinity purified on amylose depicting C-terminal cleavage activity ("MB" product). Figure 10b is a photograph of a Western immunoblotting analysis depicting C-terminal cleavage activity ("MB" product) using anti-MBP antibodies (anti-M). Figure 10c is an autoradiograph of an SDS-PAGE separation of the *in-vitro* translated, [^{35}S]-methionine–labeled chimera depicting C-terminal cleavage activity ("MB" product). M – MBP specific fragment, MB – MBP-BIL specific fragment, MBC – intact chimera.

FIGs. 11a-b are electrophoretic analyses depicting N-terminal auto-cleavage by *in-vivo* expressed MBP-4825rhosp-CBD chimera. Figure 11a is a photograph of a Coomassie blue stained electrophoretic separation of chimera protein products depicting N-terminal cleavage activity of protein products affinity purified on chitin ("BC" product) and affinity purified on amylose ("M" product). Figure 11b is a photograph of a Western immunoblotting analysis depicting N-terminal cleavage activity using anti-CBD antibody as a probe ("BC" product) and anti-MBP antibody as a probe ("M" product). BC – BIL-CBD specific fragment, M – MBP specific fragment, MB – MBP-BIL specific fragment, MBC – intact chimera.

FIGs. 12a-c are electrophoretic analyses depicting auto-processing and by *in-vivo* expressed MBP-BIL4_cloth-CBD chimera. Figure 12a is a photograph of a Western immunoblotting analysis of amylose (lane "A") or chitin (lane "C") purified protein products depicting auto-splicing activity using anti-CBD antibody as a probe ("MC" product). Figure 12b is a photograph of a Western immunoblotting analysis of amylose (lane "A") or chitin (lane "C") purified protein products depicting auto-splicing activity ("MC" product). Figure 12b also shows carboxy terminal auto-cleavage of protein products affinity purified via amylose based affinity chromatography (lane "A", "MB" product). Figure 12c is a photograph of a Coomassie blue stained electrophoretic separation of protein products isolated via amylose based (lane "A") or chitin based (lane "C") affinity chromatography depicting auto-splicing activity ("MC" species). Figure 12c also shows carboxy terminal auto-cleavage of protein products affinity purified via amylose based affinity

chromatography (lane "A", "MB" species). Note the very small amounts of the uncleaved precursor (lane "C", "MBC" species) suggesting very efficient autoprocessing activity by this chimera. M – MBP specific fragment, MB – MBP-BIL specific fragment, MBC – intact chimera.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is of chimeric autoprocessing polypeptides, polynucleotides encoding such polypeptides, and uses of such polypeptides and polynucleotides for reversibly binding proteins to specific substrates, reversibly binding specific substrates to each other, and auto-splicing amino acid sequences. Specifically, the present invention can be used to purify proteins, to affinity-select display phages, and to post-translationally ligate proteins together.

Before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details set forth in the following description or exemplified by the Examples. The invention is capable of other embodiments or of being practiced or carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein is for the purpose of description and should not be regarded as limiting.

Autoprocessing polypeptides, polypeptides having the capacity to post-translationally auto-cleave and/or auto-splice, can be used to greatly facilitate various industrially and scientifically important biochemical procedures, as described above. For example, such autoprocessing polypeptides can be used in applications which involve reversible binding of proteins to specific substrates, such as protein purification, and reversible binding of specific substrates to each other, such as affinity-selection of display phages.

Various chimeric autoprocessing polypeptides and methods of using such have been described in the prior art (reviewed in: Perler and Adam, 2000. Curr Opin Biotechnol. 11, 377-83; Paulus H., 2000. Annu Rev Biochem. 69, 447).

However, all such prior art chimeric autoprocessing polypeptides suffer from various drawbacks. As described above, these drawbacks include suboptimal activity, poor stability, insolubility, requirement for strong auxiliary nucleophiles causing undesirable modifications at carboxy termini of cleaved amino terminal fragments,

and undesirable amino terminal cleavages.

Thus, all prior art approaches have failed to provide optimal chimeric autoprocessing polypeptides for use in protein engineering.

While reducing the present invention to practice it was uncovered that the chimeric polypeptides of the present invention display efficient auto-cleavage, including auto-cleavage resulting in auto-splicing.

The chimeric polypeptides of the present invention comprise novel autoprocessing domains characterized by unique amino acid sequences, unique host protein/organism type origins, and unique natural biological capacities. Thus, the chimeric polypeptides of the present invention are highly novel and significantly enlarge and enhance the prior art spectrum of available types of chimeric autoprocessing polypeptides and their possible applications.

Thus, according to one aspect of the present invention, there are provided chimeric polypeptides having efficient auto-cleavage activity and comprising an autoprocessing segment having an amino acid sequence set forth by SEQ ID NO: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105 or 106.

Preferably, the amino acid sequence of the autoprocessing segment is set forth by SEQ ID NO: 12, 31, 76, or 77.

As used herein, the phrase "auto-cleavage activity" refers to cleavage of a polypeptide of the present invention in a region adjacent to the autoprocessing segment. Auto-cleavage occurs following exposure of the polypeptide of the present invention to suitable conditions in the absence of any other protein. Suitable auto-cleavage conditions are described hereinbelow.

Depending on the purpose, application and configuration, the polypeptides of the present invention may display different types of auto-cleavage activity.

Preferably, the auto-cleavage activity of the polypeptides of the present invention results in removal of a segment of the polypeptide adjacent to the amino terminal end or the carboxy terminal end of the autoprocessing segment.

Preferably, the segment of the polypeptide adjacent to the autoprocessing segment is an amino terminal segment or a carboxy terminal segment of the polypeptide.

According to one embodiment, the polypeptides of the present invention may comprise the autoprocessing domain as an amino terminal segment thereof. In this configuration, the polypeptides of the present invention may display removal of the segment thereof adjacent to the carboxy terminal end of the autoprocessing segment, i.e., the segment removed is the carboxy terminal segment of the polypeptide.

According to further embodiments, the polypeptides of the present invention may comprise the autoprocessing domain as a carboxy terminal segment thereof. In this configuration, the polypeptides of the present invention may display removal of the segment thereof adjacent to the amino terminal end of the autoprocessing segment, i.e., the segment removed is the amino terminal segment of the polypeptide.

According to yet further embodiments, the polypeptides of the present invention may comprise the autoprocessing domain as an internal segment of the polypeptide. In this configuration, the polypeptides of the present invention may display one or more of the following: removal of a segment adjacent to the amino terminal end of the autoprocessing segment, (i.e., the amino terminal segment of the polypeptide); removal of a segment adjacent to the carboxy terminal end thereof (i.e., the carboxy terminal segment of the polypeptide); removal of both the carboxy and amino terminal segments.

In some cases, removal of both segments may result in subsequent covalent fusion between the removed segments and, as such, auto-splicing of the polypeptide.

As used herein, the term "auto-splicing" refers to covalent bond formation between the amino acid residue forming the carboxy terminal end of a segment of the polypeptide adjacent to the amino terminal end of the autoprocessing domain and the amino acid residue forming the amino terminal end of a segment of the polypeptide adjacent to the carboxy terminal end of the autoprocessing domain.

As is illustrated in the Examples section which follows, the polypeptides of the present invention demonstrate such auto-splicing activity.

The polypeptides of the present invention may be advantageously used to post-translationally ligate essentially any protein to essentially any other protein via

formation of a covalent bond between amino acid residues forming complementary terminal ends thereof. This may be effected by using a polypeptide of the present invention comprising the proteins to be ligated in the configuration described hereinabove enabling auto-splicing to yield the desired ligation product.

Preferably, the covalent bond is a peptide bond. Alternately, as described above, the covalent bond may be an ester bond, such as the ester bond formed during auto-cleavage of prior art autoprocessing polypeptides.

Depending on the application and purpose, auto-cleavage of the polypeptide may be specifically induced under suitable conditions, preferably a specific temperature, pH, or concentration of dithiothreitol (DTT).

Preferably, the temperature is in the range of 33 to 41 °C, more preferably the temperature is in the range of 34 to 40 °C, more preferably the temperature is in the range of 35 to 39 °C, more preferably the temperature is in the range of 36 to 38 °C, more preferably the temperature is in the range of 36.5 to 37.5 °C, and most preferably the temperature is 37.0 °C.

Preferably, the pH is in the range of pH 7.8 to 8.2, more preferably the pH is in the range of pH 7.9 to 8.1, and most preferably the pH is 8.0.

Preferably, the concentration of dithiothreitol is in the range of 0.1–20 millimolar, more preferably the concentration of dithiothreitol is in the range of 0.2–10 millimolar, more preferably the concentration of dithiothreitol is in the range of 0.5–5 millimolar, and most preferably the concentration of dithiothreitol is in the range of 1.0–2.0 millimolar.

Without being bound to a paradigm, the present inventors are of the opinion that auto-cleavage activity, in addition to being governed by the amino acid sequence of the autoprocessing segment, is also influenced by the amino acid sequence of the segments adjacent to the autoprocessing segment.

For example, the present inventors are of the opinion that in configurations in which the polypeptide comprises a carboxy terminal segment adjacent to the carboxy terminal end of the autoprocessing segment, the efficiency of cleavage may be enhanced if such a carboxy terminal segment includes at its amino terminal end, an amino acid residue comprising a nucleophilic group such as a sulfhydryl group, or more preferably a hydroxyl group.

As such, in cases wherein the polypeptide of the present invention includes a carboxy terminal segment adjacent to the autoprocessing segment, the amino acid residue forming the amino terminal end of such a carboxy terminal segment is preferably cysteine, serine, or more preferably threonine. As is illustrated in the Examples section below, polypeptides of the present invention having such a carboxy terminal segment display auto-cleavage of the carboxy terminal segment, including auto-cleavage resulting in auto-splicing.

In configurations wherein the polypeptide of the present invention includes an amino terminal segment adjacent to the autoprocessing segment, the amino acid residue forming the carboxy terminal end of such an amino terminal segment is preferably serine. As is illustrated in the Examples section below, polypeptides of the present invention having such an amino terminal segment display auto-cleavage of the amino terminal segment, including auto-cleavage resulting in auto-splicing.

It is recognized in the art that certain prior art autoprocessing segments lack the capacity to auto-cleave the bond attaching a given terminal end of the autoprocessing segment to particular flanking amino acid residues (Perler and Adam, 2000. Curr Opin Biotechnol. 11, 377-83).

In particular, certain prior art autoprocessing amino acid sequences lack the capacity to auto-cleave the bond attaching their amino terminal end to a flanking serine residue. As is illustrated in the Examples section which follows, and in sharp contrast to polypeptides containing such prior art autoprocessing amino acid sequences, the polypeptides of the present invention possess the capacity to auto-cleave such a bond.

Specific amino acid sequences of autoprocessing domains comprised in the polypeptides of the present invention may be obtained by referring to Table 1, Figures 1a-b, and Figures 2a-b of the Examples section below. Table 1 provides database coordinates which can be used to retrieve the nucleic acid sequences encoding such amino acid sequences. Such amino acid sequences may easily be determined from such nucleic acid sequences by the ordinarily skilled artisan using a suitable nucleic acid-to-amino acid sequence translation software, such as, for example translation software made publicly available by the National Center for Biotechnology Information (NCBI) or the European Molecular Biology Laboratory (EMBL) on the

World Wide Web (WWW). Figures 1a-b provide specific amino acid sequences of various autoprocessing domains comprised in the polypeptides of the present invention. Figures 2a-b provide amino acid sequence motifs which define autoprocessing domains comprised in the polypeptides of the present invention.

For example, as is shown in the Examples section which follows, polypeptides of the present invention comprising autoprocessing segments FhaB_psesy (SEQ ID NO: 12), BIL4_cloth (SEQ ID NO: 31), 4825_rhosp (SEQ ID NO: 76) or BIL2_rhosp (SEQ ID NO: 77) display auto-cleavage activity, including auto-cleavage activity resulting in auto-splicing.

The polypeptides of the present invention may further comprise at least one affinity tag capable of specifically binding a substrate. As is further described hereinbelow, the affinity tag type depends on the intended use of the polypeptide.

As used herein, the phrase "affinity tag" refers to any moiety (preferably a peptide or polypeptide moiety) which is capable of specifically binding a substrate.

While the substrate can be essentially any substance or particle which can be specifically bound by the affinity tag, the substrate is preferably a molecule, a compound, a virus, or a cell.

The polypeptides of the present invention may comprise essentially any affinity tag.

Examples of peptide/polypeptide affinity tags include streptavidin, His-tags, strep-tags, epitope tags, maltose-binding proteins, and chitin-binding domains.

His-tags (histidine tags) consist of a chain of 2 to 10, most preferably 6, contiguous histidine amino acid residues. His-tags have the capacity to specifically bind substrates including nickel. Ample guidance regarding tagging polypeptides with His-tags is available in the literature of the art (for example, refer to: Sheibani N. 1999. Prep Biochem Biotechnol. 29:77). Purification of molecules comprising histidine tags is routinely effected using nickel-based automatic affinity column purification techniques. A suitable capture ligand for histidine-tagged molecules is the anti histidine tag single chain antibody 3D5 (Kaufmann, M. *et al.*, 2002. J Mol Biol. 318. 135-47).

Examples of epitope tags include an 11-mer *Herpes simplex* virus glycoprotein D peptide, and an 11-mer N-terminal bacteriophage t7 peptide, being commercially

known as HSVTag and t7Tag, respectively (Novagen, Madison, WI, USA), and 10- or 9-amino acid c-myc or *Hemophilus influenza* hemagglutinin (HA) peptides, which are recognized by the variable regions of monoclonal antibodies 9E10 and 12Ca5, respectively.

Strep-tags are peptides having the capacity to specifically bind streptavidin. Ample guidance regarding the use of strep-tags is provided in the literature of the art (see, for example: Schmidt, TGM. and Skerra, A. 1993. Protein Eng. 6, 109; Schmidt TGM. *et al.*, 1996. Journal of Molecular Biology 255, 753-766; Skerra A. and Schmidt TGM., 1999. Biomolecular Engineering 16, 79-86; Sano T. and Cantor CR. 2000. Methods Enzymol. 326, 305-11; Sano T. *et al.*, 1998. Journal of Chromatography B 715, 85-91).

Preferably, the affinity tag is a maltose-binding domain or a chitin-binding domain.

Preferably, the maltose-binding domain is *malE*-encoded maltose-binding protein (MBP). Ample guidance regarding the use of maltose-binding protein as an affinity tag is provided in the Examples section which follows and in the literature of the art (see, for example: Guan M. *et al.*, 2002. Protein Expr Purif. 26, 229-34; Cattoli F. and Sarti GC., 2002. Biotechnol Prog. 18, 94-100).

In cases where the affinity tag is a maltose-binding protein, the substrate is preferably amylose, a specific ligand of such an affinity tag. Alternately, the substrate may be maltose, also a specific ligand of such an affinity tag.

As is shown in the Examples section below, polypeptides of the present invention comprising maltose-binding protein (MBP) can specifically bind a support including amylose.

Preferably, the chitin-binding domain is *B. circulans cbd*-encoded chitin binding domain (CBD). Ample guidance regarding the use of chitin-binding domain as an affinity tag is provided in the Examples section which follows and in the literature of the art (see, for example: Humphries HE. *et al.*, 2002. Protein Expr Purif. 26, 243-8; Chong S. *et al.*, 1997. Gene 192, 271-81).

In cases where the affinity tag is *cbd*-encoded chitin binding domain (CBD), the substrate is preferably chitin, a specific ligand of such an affinity tag.

As is illustrated in the Examples section which follows, polypeptides of the

present invention comprising *cbd*-encoded chitin-binding domain can specifically bind a support including chitin.

The polypeptides of the present invention can be generated using chemical synthesis approaches or preferably recombinant techniques.

While reducing the present invention to practice, nucleic acid sequences encoding polypeptides having putative autoprocessing segments were identified in nucleic acid sequence databases (see Examples section below for further detail). The sequences were analyzed and the autoprocessing segment encoding regions were identified and used to generate polynucleotides encoding the polypeptides of the present invention.

Thus, according to another aspect of the present invention there is provided a polynucleotide sequence which encodes the auto-cleavable polypeptide of the present invention.

The polynucleotides of the present invention can be assembled from genomic, and/or complementary sequences.

As used herein, the phrase "complementary sequence" refers to a polynucleotide having a nucleic acid sequence resulting from reverse transcription of messenger RNA using a reverse transcriptase or any other RNA dependent DNA polymerase. Such sequences can be subsequently amplified *in-vivo* or *in-vitro* using a DNA dependent DNA polymerase.

As used herein, the phrase "genomic sequence" refers to a polynucleotide derived from a chromosome which thus reflects a contiguous portion of a chromosome.

In the case of a polynucleotide of the present invention encoding an autoprocessing domain expressed in a prokaryotic organism, the nucleic acid sequence encoding the autoprocessing domain may be conveniently generated via polymerase chain reaction (PCR) amplification using genomic DNA of the prokaryotic organism as a template.

Alternately, in the case of a polynucleotide of the present invention encoding an autoprocessing domain expressed in a eukaryotic organism, the nucleic acid sequence encoding the autoprocessing domain may be conveniently generated via PCR amplification using a cDNA library derived from the organism as a template.

Suitable oligonucleotide primers for PCR amplifying nucleic acid sequences encoding specific autoprocessing domains comprised in polypeptides of the present invention can be designed using the nucleic acid sequences identified by the database coordinates provided in Table 1 of the Examples section which follows.

For example, as is illustrated in the Examples section below, oligonucleotide primers suitable for PCR amplifying nucleic acid sequences encoding the autoprocessing domains FhaB_psesy, or BIL2_rhosp were derived from nucleic acid sequences retrieved using the relevant database coordinates provided in Table 1 of the Examples section below. The nucleic acid sequences of such primers are set forth by SEQ ID NOs: 4–5, or 6–7, respectively. Ample guidance for determining suitable reaction conditions for amplifying nucleic acid sequences encoding the aforementioned autoprocessing domains is provided in the Examples section which follows. PCR amplification of nucleic acid sequences is a commonly performed procedure and suitable primers and reaction conditions for a broad range of such procedures can generally be routinely determined by one of ordinary skill in the art, for example via suitable software, such as, for example, OLIGO 4.0 (National Biosciences, Plymouth, Minn.).

It will be appreciated that since autoprocessing activity is a characteristic of the amino acid sequence, one of ordinary skill in the art may alternatively use the amino acid sequences provided herein as a template for designing nucleic acid sequences which encode such amino acid sequences, and which take into consideration parameters such as codon usage which may increase the efficiency of expression of such sequences in specific organisms.

As described above, the polypeptides of the present invention may further comprise at least one affinity tag. In order to produce polypeptides of the present invention comprising affinity tags, coding nucleotides are formed comprising nucleic acid sequences encoding such affinity tags.

As described hereinabove, methods of generating nucleic acid sequences encoding affinity tags are well known to one of ordinary skill in the art. For example, nucleic acid sequences encoding affinity tags can be advantageously generated by PCR amplification of nucleic acid sequences encoding such affinity tags.

For example, as described and demonstrated in the Examples section which

follows, suitable oligonucleotide primers for amplification of nucleic acid sequences encoding *B. circulans cbd*-encoded chitin-binding domain are set forth in SEQ ID NOs: 1–2.

Alternately, autoprocessing domains may be conveniently cloned into nucleic acid constructs configured for expressing fusion proteins comprising an affinity tag fused to a polypeptide encoded by a nucleic acid insert cloned into such a construct.

As is shown in the Examples section below, polynucleotides of the present invention comprising nucleic acid sequences encoding the affinity tag maltose-binding protein were assembled by cloning a nucleic acid sequence encoding an autoprocessing domain of the polypeptides of the present invention into the expression construct pMALC2 (New England Biolabs) which is designed to express fusion proteins comprising maltose-binding protein fused to a polypeptide encoded by a cloned insert.

Such nucleic acid constructs can be advantageously used to insert and/or express a chimeric polynucleotide within a host cell.

Thus, according to yet another aspect of the present invention there is provided a nucleic acid construct comprising a polynucleotide of the present invention.

The nucleic acid constructs of the present invention preferably comprise suitable promoter sequences so as to enable efficient expression of the polynucleotide of the present invention in an expression system.

Expression of the polynucleotides of the present invention may be conveniently controlled using an inducible promoter. A suitable inducible promoter is an isopropyl beta-D-thiogalactoside (IPTG)-inducible promoter, such as a T7 promoter. As described in the Examples section which follows, a T7 promoter can be used to drive IPTG-inducible expression of a polynucleotide of the present invention in a suitable cell-free expression system or in a cellular expression system.

IPTG-induced expression of polynucleotides under the regulatory control of T7 promoters is widely practiced in the art by the ordinarily skilled practitioner and ample guidance regarding the use of such promoters is available in the literature of the art (see, for example, Sambrook *et al.*, infra).

The nucleic acid constructs of the present invention can be used to produce the polypeptides of the present invention in a suitable expression system.

Thus, according to still another aspect of the present invention there is provided a method of generating a polypeptide of the present invention.

The method is effected by generating a chimeric amino acid sequence including an autoprocessing segment of the present invention. Preferably the chimeric amino acid sequence is generated by expressing a polynucleotide of the present invention in an expression system suitable for generating the chimeric amino acid sequence from the chimeric polynucleotide.

The nucleic acid constructs of the present invention can be used to express the polypeptides of the present invention in various expression systems, including any cellular or cell-free expression systems suitable for expressing recombinant proteins such as the polypeptides of the present invention.

When used to express the polypeptides of the present invention in a cell-free expression system, the constructs of the present invention may be advantageously expressed in any suitable *in-vitro* transcription/translation system.

Numerous *in-vitro* transcription/translation systems are commercially available for expressing recombinant proteins such as the polypeptides of the present invention.

For example, a suitable cell-free expression system for expressing nucleic acid constructs of the present invention is an *E. coli* S30 extract expression system, as described and as demonstrated in the Examples section below.

Numerous cellular expression systems, including yeast, bacterial, insect, and mammalian cells can be employed to express the nucleic acid constructs of the present invention.

As described and illustrated in the Examples section which follows, the polypeptides of the present invention may be advantageously expressed in *E. coli* by transforming *E. coli* with the nucleic acid constructs.

Transformation of *E. coli* with nucleic acid constructs is a routine procedure widely practiced in the art (see, for example, Sambrook *et al.*, infra).

For example, for expression of the polypeptides of the present invention using the nucleic acid constructs of the present invention in *E. coli*, competent cells capable of DNA uptake may be prepared from cells harvested in exponential growth phase and rendered competent via the widely practiced $CaCl_2$ method. Addition of $MgCl_2$ or

RbCl to the transformation reaction medium may be employed to increase transformation efficiency. Alternative transformation methods include methods such as electroporation or host cell protoplast transformation.

As described hereinabove, the capacities of the polypeptides of the present invention to specifically bind substrates and to auto-cleave can be advantageously used in various practical applications involving reversible binding of substrates, such as protein purification.

Thus, according to a further aspect of the present invention there is provided a method of purifying a protein.

The method is effected by generating a polypeptide of the present invention comprising an autoprocessing segment being terminally attached to, or flanked by, an amino acid sequence of the protein, immobilizing the polypeptide to a support, and subjecting the immobilized polypeptide to suitable conditions for enabling auto-cleavage resulting in removal of the protein from the polypeptide.

According to a preferred embodiment, the polypeptide of the present invention is configured such that one terminal end of the autoprocessing segment is adjacent to a terminal segment of the polypeptide being the protein to be purified and the other terminal end of the autoprocessing segment is adjacent to a terminal segment of the polypeptide comprising an affinity tag.

According to this embodiment, the polypeptide of the present invention is preferably immobilized via a specific binding of the affinity tag to a specific ligand thereof included in the support. Optionally, in cases where auto-cleavage further results in detachment of the autoprocessing segment from the terminal segment of the polypeptide comprising the affinity tag, the method may advantageously further comprise the step of separating the protein from the autoprocessing segment so as to further facilitate purification of the protein. Such separation may be effected as described further hereinbelow.

According to another embodiment, the polypeptide of the present invention consists of a chimera comprising the autoprocessing segment fused to the protein to be purified.

According to this embodiment, the polypeptide of the present invention is preferably immobilized via a specific binding of the autoprocessing segment to a

specific ligand of the autoprocessing segment included in the support.

According to yet another embodiment, the polypeptide of the present invention consists of an autoprocessing segment being flanked at its amino terminal end with an amino terminal segment of the amino acid sequence of the protein to be purified, and being flanked at its carboxy terminal end with the carboxy terminal segment of the amino acid sequence of the protein to be purified complementing the amino terminal segment of the amino acid sequence of the protein to be purified.

According to this embodiment, the polypeptide of the present invention is preferably immobilized via a specific binding of the autoprocessing segment to a specific ligand of the autoprocessing segment included in the support, and auto-cleavage results in auto-splicing of the complementary amino and carboxy terminal segments of the protein to be purified to thereby release the protein.

In embodiments in which the polypeptide of the present invention is immobilized via the autoprocessing segment, the specific ligand included in the support is preferably an antibody or antibody fragment capable of specifically binding the autoprocessing segment.

Purification of a protein according to the method of the present invention may be advantageously effected via standard affinity chromatography techniques. For example, a suitable support for immobilizing a polypeptide of the present invention may be an affinity resin coupled to a specific ligand of the polypeptide of the present invention packed in a standard affinity purification column. Following subjecting of the support-bound polypeptide of the present invention to conditions suitable for auto-cleavage thereof, the highly purified protein released from the support-bound autoprocessing segment may be conveniently recovered as a flow-through fraction eluted from the column.

In the case described hereinabove, wherein auto-cleavage further results in detachment of the autoprocessing segment from the support-bound segment of the polypeptide of the present invention, separating the protein from the autoprocessing segment may be effected analogously to the method described hereinabove by using such standard affinity chromatography techniques wherein the affinity resin includes a specific ligand of the autoprocessing segment.

Alternately, various methods suitable for separating such mixtures of

polypeptides may be practiced by the ordinarily skilled artisan. Such techniques include, for example high-performance liquid chromatography (HPLC), size-exclusion chromatography, and similar methodologies.

Ample guidance regarding chromatographic isolation of proteins is widely available in the literature of the art (see, for example: Wilchek M. and Chaiken I., 2000. Methods Mol Biol 147, 1–6; Jack GW., 1994. Mol Biotechnol 1, 59-86; Narayanan SR., 1994. Journal of Chromatography A 658, 237-258; Nisnevitch M. and Firer MA., 2001. J Biochem Biophys Methods 49, 467-80; Janson JC. and Kristiansen T. in Packings and Stationary Phases in Chromatography Techniques. Unger KK. (ed.), Marcel Dekker, New York, pp. 747 (1990); Clonis YD: HPLC of Macromolecules: A Practical Approach, IRL Press, Oxford, pp. 157 (1989); Nilsson J. *et al.*, 1997. Protein Expr Purif. 11, 1-16).

As described in the Examples section which follows, amylose affinity ligand based column chromatography of a polypeptide of the present invention comprising the autoprocessing domain BIL2_rhosp flanked at its amino terminal end with an amino terminal segment including the affinity tag MBP, and flanked at its carboxy terminal end with the amino acid sequence of a protein to be purified resulted in column-retention of a segment of the polypeptide of the present invention lacking the amino acid sequence of the protein to be purified (Figure 10a), thereby demonstrating the utility of the method of the present invention for purifying proteins.

As described in the Examples section which follows, chitin affinity ligand based column chromatography of a polypeptide of the present invention comprising the autoprocessing domain 4825_rhosp flanked at its carboxy terminal end with an carboxy terminal segment including the affinity tag CBD, and flanked at its amino terminal end with the amino acid sequence of a protein to be purified resulted in column-retention of a segment of the polypeptide of the present invention lacking the amino acid sequence of the protein to be purified (Figure 11b), thereby demonstrating the utility of the method of the present invention for purifying proteins.

According to the teachings of the present invention, the polypeptides of the present invention may include affinity tags flanking the autoprocessing segment.

Thus, according to yet a further aspect of the present invention there is provided a method of reversibly attaching a first substrate to a second substrate.

The first and second substrates may be the same or may be different.

The method is effected using a polypeptide of the present invention in which the autoprocessing segment is flanked by a first amino acid sequence capable of binding the first substrate and a second amino acid sequence capable of binding the second substrate, such that auto-cleavage releases the first amino acid sequence from the second amino acid sequence. Exposing the first substrate and the second substrate to the polypeptide of the present invention generates a complex including the first substrate attached via the polypeptide to the second substrate. Following generation thereof, the complex is subjected to suitable conditions for auto-cleavage, thereby detaching the first substrate from the second substrate.

Complex generation may be effected in various ways, depending on the application and purpose. For example, complex generation may be effected wherein neither, one or both substrates is included in, or consists of, a support specifically binding an affinity tag comprised in the polypeptide of the present invention.

According to a preferred embodiment, the method is used for reversibly attaching a first substrate being a protein displayed by a bacteriophage to a second substrate specifically binding the phage-displayed protein, which substrate being included in a support.

According to this embodiment, the method is effected using a polypeptide of the present invention comprising a first amino acid sequence having the capacity to specifically bind the substrate and a second amino acid sequences having the capacity to bind the phage-displayed protein.

The method according to this aspect of the present invention may be advantageously employed with phage-display libraries for selecting bacteriophages displaying a protein having high affinity to a specific substrate. This may be effected by exposing a phage display library to a support including a substrate being a target molecule to which a high affinity ligand is desired. Elements of the phage display library not being bound with high affinity to the support may be washed and recovery of phages specifically binding the substrate with high affinity via a displayed protein capable of specifically binding the target molecule with high affinity may be conveniently recovered by subjecting the support-bound phages to conditions suitable for auto-cleavage of the polypeptide of the present invention so as to effect the

detachment thereof from the support.

Additional objects, advantages, and novel features of the present invention will become apparent to one ordinarily skilled in the art upon examination of the following examples, which are not intended to be limiting. Additionally, each of the various embodiments and aspects of the present invention as delineated hereinabove and as claimed in the claims section below finds experimental support in the following examples.

EXAMPLES

Reference is now made to the following examples, which together with the above descriptions, illustrate the invention in a non limiting fashion.

Generally, the nomenclature used herein and the laboratory procedures utilized in the present invention include molecular, biochemical, microbiological and recombinant DNA techniques. Such techniques are thoroughly explained in the literature. See, for example, "Molecular Cloning: A laboratory Manual" Sambrook *et al.*, (1989); "Current Protocols in Molecular Biology" Volumes I–III Ausubel, R. M., ed. (1994); Ausubel *et al.*, "Current Protocols in Molecular Biology", John Wiley and Sons, Baltimore, Maryland (1989); Perbal, "A Practical Guide to Molecular Cloning", John Wiley & Sons, New York (1988); Watson *et al.*, "Recombinant DNA", Scientific American Books, New York; Birren *et al.* (eds) "Genome Analysis: A Laboratory Manual Series", Vols. 1–4, Cold Spring Harbor Laboratory Press, New York (1998); methodologies as set forth in U.S. Pat. Nos. 4,666,828; 4,683,202; 4,801,531; 5,192,659 and 5,272,057; "Cell Biology: A Laboratory Handbook", Volumes I–III Cellis, J. E., ed. (1994); "Current Protocols in Immunology" Volumes I–III Coligan J. E., ed. (1994); Stites *et al.* (eds), "Basic and Clinical Immunology" (8th Edition), Appleton & Lange, Norwalk, CT (1994); Mishell and Shiigi (eds), "Selected Methods in Cellular Immunology", W. H. Freeman and Co., New York (1980); available immunoassays are extensively described in the patent and scientific literature, see, for example, U.S. Pat. Nos. 3,791,932; 3,839,153; 3,850,752; 3,850,578; 3,853,987; 3,867,517; 3,879,262; 3,901,654; 3,935,074; 3,984,533; 3,996,345; 4,034,074; 4,098,876; 4,879,219; 5,011,771 and 5,281,521;

"Oligonucleotide Synthesis" Gait, M. J., ed. (1984); "Nucleic Acid Hybridization" Hames, B. D., and Higgins S. J., eds. (1985); "Transcription and Translation" Hames, B. D., and Higgins S. J., eds. (1984); "Animal Cell Culture" Freshney, R. I., ed. (1986); "Immobilized Cells and Enzymes" IRL Press, (1986); "A Practical Guide to Molecular Cloning" Perbal, B., (1984) and "Methods in Enzymology" Vol. 1–317, Academic Press; "PCR Protocols: A Guide To Methods And Applications", Academic Press, San Diego, CA (1990); Marshak *et al.*, "Strategies for Protein Purification and Characterization – A Laboratory Course Manual" CSHL Press (1996); all of which are incorporated by reference as if fully set forth herein. Other general references are provided throughout this document. The procedures therein are believed to be well known in the art and are provided for the convenience of the reader.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below.

EXAMPLE 1

Bacterial-intein like (BIL) domains: novel auto-cleaving/auto-splicing protein domains

Autoprocessing polypeptides capable of auto-cleavage have been shown to be uniquely useful in a wide range of protein engineering applications, for example for protein purification without the requirement for proteases. However, all prior art autoprocessing polypeptides suffer from various drawbacks, including suboptimal activity, stability, solubility, and requirement for auxiliary molecules causing undesirable protein modifications. In order to enlarge and enhance the current repertoire of autoprocessing polypeptides, the present inventors have identified, generated and demonstrated the functionality of novel autoprocessing polypeptides, as follows.

Materials and Methods:

In order to identify novel auto-cleaving/-splicing proteins, databases storing

genomic sequences of various organisms, including bacterial pathogens were searched for open reading frames (ORFs) coding for protein sequences containing Hint domains. Following identification of protein sequences containing Hint domains, such protein sequences were cloned and tested for auto-cleaving/-splicing activity, as described below.

Data sources: BILs were identified in bacterial genomes by searching the following databases: National Center for Biotechnology Information (NCBI) sequence databases for *Brucella melitensis* (*B. melitensis*) 16M, *Streptomyces coelicolor* (*S. coelicolor*) A3(2), *Neisseria meningitidis* (*N. meningitidis*) MC58, *N. meningitidis* Z2491, *Pseudomonas fluorescens* (*P. fluorescens*) PfO-1, *Leptospira interrogans* (*L. interrogans*) *56601*, *Streptomyces avermitilis* (*S. avermitilis*) MA-468, *Pirellula species* 1, and *Chromobacterium violaceum* (*C. violaceum*) *ATCC 12472* sequences; Integrated Genomics (http://www.integratedgenomics.com) sequence databases for *Rhodobacter capsulatus* (*R. capsulatus*) SB1003 genomic sequences (Haselkorn *et al.*, 2001. Photosynthesis Research 70, 43-52) and *Methylobacterium extorquens* (*M. extorquens*) AM1,; Joint Genome Institute (http://www.jgi.doe.gov) sequence databases for *Rhodobacter sphaeroides* (*R. sphaeroides*) 2.4.1 (Mackenzie *et al.*, 2001. Proc Natl Acad Sci U S A. 99, 2275-2280), *Magnetospirillum magnetotacticum* (*M. magnetotacticum)* MS-1, *Clostridium thermocellum* (*C. thermocellum*) ATCC 27405, and *Thermobifida fusca* (*T. fusca*) YX genomic sequences; The Institute for Genomic Research (http://www.tigr.org) sequence databases for *Pseudomonas syringae* (*P. syringae*) DC3000 (Fouts *et al.*, 2002. Proc Natl Acad Sci U S A 99, 2275-2280), *Silicibacter pomeroyi* (*S. pomeroyi*) DSS-3, *Gemmata obscuriglobus* (*G. obscuriglobus*) UQM 2246, *Myxococcus xanthus* (*M. xanthus*) DK1622, and *Verrucomicrobium spinosum* (*V. spinosum*) DSM 4136 sequences; The Sanger Institute (http://www.sanger.ac.uk) sequence databases for *Neisseria meningitidis* (*N. meningitidis*) FAM18, *P. fluorescens* SBW25, and *Rhizobium leguminosarum* (*R. leguminosarum*) bv. viciae 3841 sequences; University of Oklahoma, Advanced Center for Genome Technology (http://www.genome.ou.edu) sequence databases for *Neisseria gonorrhoeae* (*N. gonorrhoeae*) FA1090 genomic sequences; Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu) sequence databases for *Mannheimia haemolytica* (*M. haemolytica*) PHL213 genomic

sequences; and Kazusa DNA Research Institute database (Japan; http://www.kazusa.or.jp) for *Gloeobacter violaceus* (*G. violaceus*) PCC 7421 sequences.

BIL domain nomenclature: BIL domains were named using the format "a_b" where "b" is an abbreviation of the bacterial species of origin and where "a" is either host protein name (e.g., "FhaB" or "MafB"); an arbitrary BIL# designation; an Integratedgenomics Database (http://ergo.integratedgenomics.com/R_capsulatus.html) number for *R. capsulatus*; a Computational Biology Program at ORNL (http://genome.ornl.gov/microbial/rsph) analysis code for *R. sphaeroides*; or a gene number for *B.melitensis* strain 16M and *N.meningitidis* strains MC58 and Z2491. Available gene identifier accession numbers and further information relevant to identified BIL domains are provided in Table 1, below.

Computational sequence analysis: The BLAST software package of the NCBI was used for sequence-to-sequence searches (Altschul, SF. *et al.*, 1997. Nucleic Acids Res. 25, 3389) of BIL domains with BIL domains and with intein sequences, and the BLIMPS software was employed for block-to-sequence searches (Henikoff S. *et al.*, 1995. Gene 163, GC17). Multiple block sequence alignments were constructed using BLOCKMAKER (Henikoff S. *et al.*, 1995. Gene 163, GC17) and MACAW (Schuler GD. *et al.*, 1991. Structure, Function and Genetics 9, 180) software, as previously described (Pietrokovski S., 1998. Protein Science 7, 64). BIL domains were aligned with other BIL domains having higher scores than with intein sequences and alignments of BIL domains with each other was across their whole, or almost whole, lengths (results not shown). This is also a practical way to classify BIL domains as such. Phylogenetic analysis was performed using the PHYLIP software package (Felsenstein J., 1989. Cladistics 5, 164) version 3.55.

Generation of BIL domain phylogeny dendrograms: BIL domain phylogeny dendrograms were computed from DNA multiple sequence alignment of 49 mostly complete BIL domains aligned across 201 positions, coding for 67 amino acids which could be confidently aligned across BIL domains. Nodes with bootstrap values below 440/1000 were collapsed, and bootstrap values above 800/1000 are shown. Bootstrap values of the nodes grouping all A-type and B-type BILs are 441 and 519, respectively. The *D. melanogaster* Hedgehog Hint domain (Porter JA., *et al.*, 1996.

Cell 86, 21) was used as an outgroup to root the tree. The dendrogram was calculated using the DNADIST program (version 3.5) of the PHYLIP software package (Felsenstein J., 1989. Cladistics 5, 164). Results were verified against those obtained using CLUSTALW software (Thompson JD. *et al.*, 1994. Nucl Acid Res. 22, 4673) and from protein multiple sequence alignments obtained using PHYLIP, PROTDIST, and CLUSTALW software.

BIL functional activity assays:

In order to analyze the capacity of BIL domains to auto-cleave/auto-splice flanking sequences, genetic sequences encoding BIL domains and portions of flanking sequences were cloned for expression as chimeric proteins tagged at their amino terminal ends with the *malE* gene-encoded maltose-binding protein (MBP) affinity tag, and at the carboxy terminal end with the *B. circulans cbd* gene-encoded chitin-binding domain (CBD) affinity tag. These chimeras were expressed in an *in-vitro* transcription/translation system, or overexpressed *in-vivo* in *E. coli,* and resulting protein products were analyzed for evidence of BIL domain-mediated autoprocessing activity.

Such chimeras were cloned using BIL domain genetic sequences encoding:

(i) the Type A BIL domain FhaB_psesy (Table 1, Figure 1a) and its downstream-flanking threonine residue to generate the chimera "MBP-PsyBIL-CBD"; and

(ii) the Type B BIL domain BIL2_rhosp (Table 1, Figure 1b) including 32 amino terminal-flanking and 11 carboxy terminal-flanking amino acids to generate the chimera "MBP-RspBIL2-CBD".

Constructs: *B. circulans cbd* gene sequences encoding CBD were cloned by PCR from expression vector pTYB2 (New England Biolabs, Beverly, MA) using the primers 5′-AAATGTCGACTGCGGTGGCCTGACC-3′ (SEQ ID NO: 1) and 5′-TGTCGTATTGCTTCCTTTCGGGCTT-3′ (SEQ ID NO: 2), and inserted, including the upstream linker 5′-TGCGGTGGCCTGACCGGTCTGAACTCAGGCCTC-3′ (SEQ ID NO: 3), into the SalI/PstI linearized, isopropyl beta-D-thiogalactoside (IPTG)-inducible MBP-fusion protein expression construct pMALC2 to generate the MBP-CBD fusion protein expression construct pC2C. Construct pC2C was used in functional assays as a positive control for expression of MBP-CBD.

For MBP-PsyBIL-CBD expression, genetic sequences encoding PsyBIL and flanking sequences were PCR amplified from *P. syringae* DC3000 strain genomic DNA (kindly provided by Dr. G. Sessa, Tel-Aviv University) using the primers 5′-AAAAGGATCCTGCTTTGCGGCCGGAACGA-3′ (SEQ ID NO: 4) and 5′-AAAATCTAGAGGTATTATGCACCCATGTCTTG-3′ (SEQ ID NO: 5), and cloned in BamHI/XbaI linearized pC2C, between the *malE* MBP-encoding sequences and the CBD expressing *cbd* sequences to generate the MBP-PsyBIL-CBD expression construct pC2C-PsyBIL.

For MBP-RspBIL2-CBD expression, genetic sequences encoding RspBIL2 and flanking sequences were PCR-amplified from *R. sphaeroides* 2.4.1 strain genomic DNA (supplied by Dr. Steven L. Porter, Department of Biochemistry, University of Oxford) using the primers 5′-GAATTCGGTGATTCATCCTTGGGGCGA-3′ (SEQ ID NO: 6) and 5′-TCTAGAAAAACACGGCAAGGGCGAGCGG-3′ (SEQ ID NO: 7), and cloned in EcoRI/XbaI linearized pC2C, between the MBP-encoding *malE* sequences and the CBD-encoding *cbd* sequences to generate the MBP-RspBIL2-CBD expression construct pC2C-RspBIL2.

the MBP-BIL4cloth-CBD expression construct pC2C-BIL4cloth.

Polymerase chain reactions were performed using a Biometra thermal cycler in a 50 µl reaction mixture containing Taq polymerase buffer (Sigma, St. Louis, MI), 1 µl Taq DNA polymerase, 200 mM dNTP, 10 mM of each primer and 100 ng genomic DNA

The chimeras were constructed such that carboxy terminal or amino terminal cleavage thereof was expected to generate [MBP-BIL + CBD] or [MBP + BIL-CBD] specific protein products, respectively. The MBP- or CBD-containing protein products were expected to vary in size according to the size of the BIL domain flanking sequences included in the BIL sequences cloned in the chimeric proteins. Auto-splicing by the chimeras was expected to generate MBP-CBD protein products having a molecular weight varying according to the size of the BIL domain flanking sequences included in the cloned BIL segment.

Chimeras were expressed and analyzed for evidence of BIL processing activity, as described below.

In-vitro BIL protein expression and activity assays: *In-vitro* transcription-translation of MBP-BIL-CBD and MBP-CBD chimeric proteins was achieved using chimera expression constructs pC2C-RspBIL2 and pC2C-RspBIL2a, and expression construct pC2C as DNA templates, respectively, using *E. coli* S30 extract for circular DNA system (Promega Kit #L1030, Promega, Madison WI). Reactions were carried out according to the manufacturer's instructions, using a reaction containing 0.25 mM [^{35}S]-methionine, 220 nmol of expression construct DNA as template, 1–2 mM dithiothreitol, and having a pH of about 8.0. Reactions were incubated at 37 °C for 90–120 minutes. Prior to electrophoresis of expressed protein, 5 µl or 10 µl aliquots of reaction mixtures were mixed with four volumes of acetone in order to remove polyethylene glycol. Acetone precipitation was followed by centrifugation at 12,000 × g for 5 minutes. The supernatant was discarded and the protein-containing pellet was mixed with gel loading buffer to give a final concentration of 0.06 M Tris-Cl, 2 % SDS, 10 % (v/v) glycerol, and 0.01 % bromophenol blue. Proteins were separated via 7.5 % or 10 % SDS-PAGE, and the separated proteins were visualized using a phosphor-imaging screen. Phosphor-imaging signals were quantified using NIH IMAGE 1.62 software. Product quantities were derived from values of three independent experiments averaged for each sample together with their standard deviation of the means. The molar percentage of each product was calculated.

The expected molecular weights of MBP-RspBIL2-CBD, and of its splicing product MBP-CBD, its carboxy terminal cleavage product MBP-RspBIL2, and its MBP-containing amino terminal cleavage product are 68.0, 55.1, 60.5 and 46.7 kDa, respectively.

The expected molecular weights of control vector pC2C-expressed MBP-CBD and of its MBP portion were 50.3 and 43.0 kDa, respectively.

In-vivo Type A BIL domain protein expression, purification and activity assay: Competent TB1 *E. coli* cells (NEB, Beverly, MA) were transformed with constructs for expression of MBP-BIL-CBD chimeras. Transformants were plated on LB agar supplemented with ampicillin (100 µg/ml). Single colonies were used to inoculate 3 ml aliquots of LB medium supplemented with ampicillin (100 µg/ml). Following incubation at 37 °C for 16 hours with shaking, 1 ml of culture was used to inoculate a 2 liter flask containing 500 ml of LB supplemented with ampicillin (100

μg/ml). Incubation was continued at 37 °C with shaking until the optical density at 600 nm was 0.6, at which point IPTG was added to a final concentration of 0.3 mM. After further incubation for 3 hours, cells were harvested by centrifugation at 5,000 × *g* for 20 minutes, re-suspended in solution containing 20 mM Tris (pH 7.4), 200 mM NaCl, and protease inhibitor cocktail (Sigma, St. Louis, MI), and lysed by sonication. Lysates were centrifuged at 17,000 × *g* for 20 minutes to remove cell debris, and supernatants were harvested for subsequent analyses. Proteins were then affinity purified with either chitin (NEB, Beverly, MA) or amylose beads (NEB, Beverly, MA) which bind to the CBD or MBP affinity tags included in the chimeric protein. Elution of protein from beads prior to electrophoresis was performed by mixing the protein-bound beads with SDS-PAGE sample loading buffer.

Western immunoblotting assays and protein staining: Products generated by expression of MBP-BIL-CBD chimeras were separated by SDS-PAGE. Briefly, protein samples were mixed with protein loading buffer to give a final concentration of 0.06 M Tris-Cl, 2 % SDS, 10 % (v/v) glycerol, 0.1 M dithiothreitol, and 0.01 % bromophenol blue. All samples were boiled for 3 minutes prior to electrophoresis. Separated proteins were analyzed by Western immunoblotting using either monoclonal mouse anti-MBP (Novus Biologicals, Inc. Littleton, CO) antibody for identification of the MBP tag, or polyclonal rabbit anti-CBD (NEB, Beverly, MA) for identification of the CBD tag. Secondary antibodies used were HRP conjugated goat anti-mouse IgG or goat anti-rabbit IgG (Jackson ImmunoResearch Laboratories Inc., West Grove, PA). Relative apparent molecular weights were calculated using TriChromoRanger (Pierce, Rockford IL) prestained markers.

Electrophoretic gels containing separated proteins were fixed in 40 % methanol/7 % acetic acid and stained with PhastGel Blue R stain (Pharmacia Biotech AB, Sweden). Gels were destained in 40 % methanol/7 % acetic acid and then in deionized water, and visualized protein bands were excised and electroeluted for MALDI mass spectroscopy (MS) analysis.

Electroelution of protein: Protein was electroeluted from gels at 150 volts for 2 hours in GeBAflex tubes (Gene Bio Application Ltd., Israel) using elution buffer containing 0.025 % SDS, Tris and Tricine (pH 8.5). Following electroelution SDS was removed from electroeluted protein using cold TCA:acetone precipitation in the

presence of 0.5 % sodium deoxycholate (NaDOC; T. Mehlman and A. Shainskaya, unpublished).

In-gel proteolysis: Protein bands from PhastGel Blue R stained gels were destained using multiple washes in 50 % acetonitrile in 50 mM ammonium bicarbonate. Destained protein bands were subsequently reduced, alkylated and in-gel proteolysed using either bovine trypsin (sequencing grade, Roche Diagnostics, Germany) or chymotrypsin, (Boehringer Mannheim, Germany) by incubation with 12.5 ng/µl protease in 50 mM ammonium bicarbonate at 37 °C, as previously described (Shevchenko *et al.*, 1996. Analytical Chemistry 68, 850). Extracted peptide solutions were dried for subsequent MALDI-MS analysis.

Mass Spectrometry: Intact molecular mass measurement and peptide mass mapping were performed using a Bruker Reflex III MALDI time-of-flight (TOF) mass spectrometer (Bruker, Bremen, Germany) equipped with SCOUT source, delayed ion extraction, reflector and a 337 nm nitrogen laser. Each mass spectrum was generated using data accumulated from 200 laser shots. Both external and nearby calibrations for proteins were performed using BSA and myoglobin (Sigma). For peptide mapping, internal calibration with molecular ions of regularly occurring matrix ions and peptides derived from trypsin was additionally performed to consolidate further peptide assignment.

Intact molecular weight measurements by MALDI MS: Gel electroeluted proteins were further purified by cold acetone precipitation. The dried extract from one lane of the gel was re-dissolved in 0.5 ml of 80 % formic acid and immediately diluted with water to yield a solution containing 20 % formic acid, and 50 % of this solution was applied to a target plate.

Peptide mass mapping by MALDI mass spectrometry: Aliquots of one tenth of the extracted peptide mixture volume, dissolved in 0.1 % TFA or formic acid/isopropanol/water (1:3:2), were used for MALDI-MS using the fast evaporation or dry droplet method. Matrix surfaces of α-cyano-4-hydroxycinnamic acid (4-HCCA) or 2,5-dihydroxybenzoic acid (DHB) were utilized for the fast evaporation (Jensen ON. *et al.*, 1996. Rapid Commun in Mass Spectrom. 10, 1371; Vorm O. *et al.*, 1994. Analyt Chem. 66, 3281) or dry droplet method (Kussmann K. *et al.*, 1997. J Mass Spectrom. 32, 593), respectively.

Experimental Results:

Identification of two novel bacterial intein-like domains containing Hint-like motifs: Searches of sequence databases of diverse bacterial species for Hint-like motif-containing putative ORFs identified open reading frames coding for proteins comprising two related types of novel intein-like protein domains termed by the present inventors Type A and Type B bacterial intein-like (BIL) domains. Novel type A and B BIL domain sequences identified are shown aligned in Figures 1a (SEQ ID NOs: 8-62) and 1b (SEQ ID NOs: 63-104), respectively. The bioinformatic sources used to identify these BIL domains are shown in Table 1.

An amino acid sequence motif (SEQ ID NO: 105) was identified (Figure 2a) which exclusively defines a subset of Type A BIL domains, including domains: 39_9_thefus, SCP1.201_strco, 3875_87_magma, B0372+_neimeB, B0655+_neimeB, A2115_neime, MafB1_neimeC, BIL2_neimeC, BIL4_neimeC, BIL5_neimeC, BIL6_neimeC, MafB1_neigo, BIL2_neigo, BIL3_neigo, MafB2_neigo, BIL5_neigo, BIL6_neigo, FhaB_psesy, FhaB_manha, FhaB1_psefl-PfO-1, FhaB1_psefl-SBW25, BIL6_cloth, BIL5_cloth, BIL2_cloth, BIL4_cloth, BIL1_cloth, BIL8_cloth, BIL9_cloth, BIL1_gemob, BIL2_gemob, 0709_lepin, 3725_lepin, o1078_myxxa, o1070_myxxa, BIL1_strav, BIL2_strav, BIL3_strav, BIL1_pirsp, BIL1_chrvi, o3395_versp, o5687_versp, o649_versp, BIL1_glovi, BIL2_glovi, BIL3_glovi, and BIL4_glovi.

An amino acid sequence motif (SEQ ID NO: 106) was identified (Figure 2b) which exclusively defines a subset of Type B BIL domains, including: 4825_rhosp, BIL2_rhosp, 00588_rhoca, 02710_rhoca, 01524_rhoca, 01523_rhoca, 00126_rhoca, 01216_rhoca, 00949_rhoca, 01374_rhoca, 00459_rhoca, 00460_rhoca, 00746_rhoca, 03530_rhoca, 00199_rhoca, BIL3_magma, BIL4_magma, BIL1_brusu, BIL1_unknwn, BIL2_unknwn, 06786_metex, BIL1_silpo, BIL2_silpo, BIL3_silpo, BIL4_silpo, BIL5_silpo, BIL6_silpo, BIL7_silpo, BIL8_silpo, BIL9_silpo, BIL10_silpo, BIL11_silpo, BIL12_silpo, BIL13_silpo, BIL14_silpo, BIL15_silpo, BIL16_silpo, BIL1_rhile, and II0519_brume.

This new type of domain appears in non-conserved regions of hyper-variable proteins. Thus, these domains are distinct from the Hint domains of inteins and Hog-proteins by the species and proteins in which they appear. An analysis of amino acid

residue conservation within Hint-like motifs of BIL domains and within homologous Hint domain motifs of Hog proteins and inteins is shown in (Figures 3a-z). Examination of BLAST sequence alignments (Altschul, SF. *et al.*, 1997. Nucleic Acids Result. 25, 3389) of BIL domains with BIL domains and with intein sequences showed that BIL domains aligned with each other with higher scores than with intein sequences across their whole, or almost whole, lengths (results not shown). Therefore, BIL domains were found to be distinct from inteins by their global sequence features.

Table 1. Databases used to identify BIL domains.

BIL Type	BIL domain name*	Source**	Date***	Contig/Entry	Coordinates
A	BIL1_cloth	NCBI		23022619	-
A	BIL2_cloth	NCBI		28N'aa+23020813+ 59aa+23020812	-
A	BIL3_cloth	NCBI		23022239+14N'aa	-
A	BIL4_cloth	NCBI		23020817+5N'amin o acid/gi|23020817	311-445
A	BIL5_cloth	NCBI		23020815+13N'aa	-
A	BIL6_cloth	NCBI		23022237	-
A	BIL7_cloth	NCBI		23022587+7N'aa	-
A	BIL9_cloth	NCBI		23022893+59aa+ 23022892	-
A	BIL10_cloth	NCBI		22262017	34594-34986
A	BIL11_cloth	NCBI		22262176	416-728
A	3875_87_magma	NCBI		21614488	76532-75165
A	FhaB_manha	BCM	4Oct01	C78-C85	11046-20977
A	BIL2_neigo	OU-ACGT	26Sep00	AE004969	1563413-1564129
A	BIL3_neigo	OU-ACGT	26Sep00	AE004969	1565033-1565809
A	BIL5_neigo	OU-ACGT	26Sep00	AE004969	1351509-1350766
A	BIL6_neigo	OU-ACGT	26Sep00	AE004969	1349978-1349310
A	MafB1_neigo	OU-ACGT	26Sep00	AE004969	1560214-1561941
A	MafB2_neigo	OU-ACGT	26Sep00	AE004969	1355876-1354062
A	B0369+_neimeB	NCBI		7225591 +34 N' aa	-
A	B0372+_neimeB	NCBI		7225594 +11 N' aa	-
A	B0655+_neimeB	NCBI		7225882 +14 N' aa	-
A	BIL2_neimeC	Sanger	15May02	NmC	1836857-1837573

A	BIL3_neimeC	Sanger	15May02	NmC	1838418-1838981
A	BIL4_neimeC	Sanger	15May02	NmC	1839771-1840439
A	BIL5_neimeC	Sanger	15May02	NmC	627204-627920
A	BIL6_neimeC	Sanger	15May02	NmC	628395-629102
A	MafB1_neimeC	Sanger	15May02	NmC	1833717-1835480
A	FhaB1_psefl-PfO-1	NCBI-TIGR		205922-575	3-6313
A	FhaB1_psefl-SBW25	Sanger	2Sep02	Pflu552a01	41728-29387
A	FhaB_psesy	TIGR	30Aug02	5668	5148986-5149429
A	SCP1.201_strco	NCBI		13620683 +32 N' aa	
A	39_9_thefus	JGI	1Nov00	39	13655-15508
A	BIL1_gemob	TIGR	23Sep02	14	32-20392
A	BIL2_gemob	TIGR	14Feb03	354	16692-11734
A	0709_lepin	NCBI	10Nov02	24213409	-
A	3725_lepin	NCBI	10Nov02	24216424	-
A	3719_lepin	NCBI	10Nov02	24197710	-
A	o665_myxxa	TIGR	23Apr03	168	495-3704
A	o1078_myxxa	TIGR	23Apr03	157	103477-106710
A	o1070_myxxa	TIGR	23Apr03	168	495-3704
A	BIL1_strav	NCBI	20Jul03	29826740	-
A	BIL2_strav	NCBI	20Jul03	29826826	-
A	BIL3_strav	NCBI	20Jul03	29831835	-
A	BIL1_pirsp	NCBI	20Jul03	32470666	-
A	BIL1_chrvi	NCBI	07Sep03	34104178	-
A	BIL1_glovi	Kazusa	04Sep03	gll0211	-
A	BIL2_glovi	Kazusa	04Sep03	gll0207	-
A	BIL3_glovi	Kazusa	04Sep03	gll0213	-
A	BIL4_glovi	Kazusa	04Sep03	gll0212	-
A	BIL5_glovi	Kazusa	04Sep03	gll0205	-
A	BIL6_glovi	Kazusa	04Sep03	gll0208	-
A	BIL7_glovi	Kazusa	04Sep03	gsl3615	-
A	o649_versp	TIGR	02Sep03	65738	3-1949
A	o5687_versp	TIGR	02Sep03	65921	18348-1288
A	o3395_versp	TIGR	02Sep03	65925	56853-46669
B	II0519_brume[1]	NCBI		17988864	
B	BIL2_magma	NCBI		21613062	922-1590
B	BIL3_magma	NCBI		21614112	2449-1475
B	BIL4_magma	NCBI		21614173	2216-3187
B	BIL5_magma	NCBI		21612572	3-338
B	BIL6_magma	NCBI		21613847	2033-1774
B	06786_metex	IG	Jun02	1507	6076-7113
B	00126_rhoca	IG	Dec01	2G06-2D11	114767-113670
B	00199_rhoca	IG	Dec01	2G06-2D11	178648-177806
B	00459_rhoca	IG	Dec01	2G06-2D11	434469-435083
B	00460_rhoca	IG	Dec01	2G06-2D11	435094-436191
B	00746_rhoca	IG	Dec01	2D10-2D06	2243-3079
B	00949_rhoca	IG	Dec01.	2A12-2D05	325707-326651

B	01216_rhoca	IG	Dec01	2A12-2D05	222590-223555
B	01374_rhoca	IG	Dec01	2A12-2D05	148470-149462
B	01523_rhoca	IG	Dec01	1A01-1C09	279638-280444
B	01524_rhoca	IG	Dec01	1A01-1C09	280569-281423
B	02710_rhoca	IG	Dec01	1D09-1F02	197288-199588
B	03530_rhoca	IG	Dec01	1A01-1C09	521700-521173
B	4825_rhosp	JGI	26Mar01	184	67785-67165
B	BIL2_rhosp	JGI	26Mar01	177	9673-10194
B	BIL1_silpo	TIGR	18Jun02	50	10679-14440
B	BIL2_silpo	TIGR	18Jun02	50	1-2688
B	BIL3_silpo	TIGR	18Jun02	4	10260-12356
B	BIL4_silpo	TIGR	18Jun02	290	18579-19640
B	BIL5_silpo	TIGR	18Jun02	11	3224-2217
B	BIL6_silpo	TIGR	18Jun02	199	13687-15303
B	BIL7_silpo	TIGR	18Jun02	199	15104-16399
B	BIL8_silpo	TIGR	18Jun02	32	1329-2600
B	BIL9_silpo	TIGR	18Jun02	32	28672-26588
B	BIL10_silpo	TIGR	18Jun02	60	19338-18265
B	BIL11_silpo	TIGR	18Jun02	89	10027-11181
B	BIL12_silpo	TIGR	18Jun02	126	9020-10066
B	BIL13_silpo	TIGR	18Jun02	110	13640-14440
B	BIL14_silpo	TIGR	18Jun02	125	9769-10353
B	BIL15_silpo	TIGR	18Jun02	129	7727-8257
B	BIL16_silpo	TIGR	18Jun02	195	7271-8050
B	Bil1_rhile	Sanger	14Jul03	RHIZ10E3Cb12.s1k	-
B	BIL1_unknwn	TIGR	13Mar01	14712	1124-51
B	BIL2_unknwn	TIGR	13Mar01	12703	2-1369

*the bacterial species origin of the BIL domains is identified by the bacterial species code following the underscore in the BIL name (refer to Table 2 below for the code-species correspondance).

**The sources are named as follows JGI – Joint Genome Institute (http://www.jgi.doe.gov), NCBI – (http://www.ncbi.nlm.nih.gov), Sanger – The Sanger Institute (http://www.sanger.ac.uk), OU–ACGT – University of Oklahoma, Advanced Center for Genome Technology (http://www.genome.ou.edu), TIGR – The Institute for Genomic Research (http://www.tigr.org), BCM – Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu), IG – IntegratedGenomics (http://www.integratedgenomics.com), and Kazusa – Kazusa DNA Research Institute database (Japan; http://www.kazusa.or.jp).

***Dates refer to the data release dates used.

The NCBI entries were extended as noted. Coordinates of the BIL host protein ORF are given for nucleotide contigs/entries. The positions of BILs within these ORFs are provided in Figures 1a-b.

¶An identical sequence was identified in *Brucella suis*

Table 1 cont.

Conserved motifs in BIL domains, Hog proteins and inteins: Alignments of conserved motifs in Type A BIL domains, Type B BIL domains, Hog proteins, and inteins are shown in Figures 3a-g, 3h-o, 3p-t, and 3u-z, respectively. Type A BIL domains were found to share 7 Hint-like consensus sequence motifs (Figures 3a-d and

3f-g) and one novel motif (Figure 3e) having no known Hog or intein counterpart.

Almost all Type A BIL domains were found to comprise apparent functional protein splicing active sites corresponding to those present in inteins (marked by asterisks in Figures 3u, 3w and 3z), and several are also flanked at their carboxy terminal ends with serine or threonine amino acid residues, similarly to the carboxy terminal ends of inteins.

Type A BIL domains were found to contain an invariant His-Asn amino acid residue pair adjacent to the carboxy terminal end thereof (Figure 3g), similarly to the His-Asn amino acid residue pair typically forming the carboxy terminal ends of inteins (Figure 3z, positions 7-8). Conservation of the Type A BIL domain carboxy terminal end with that of inteins suggests that, similarly to inteins, Type A BILs undergo cyclization of the Asn residue forming the carboxy terminal end thereof. However, the residue at the carboxy terminal end of Type A BIL domains (Figure 3g, position 8), which corresponds to the residue forming the amino terminal end of polypeptide segments flanking the carboxy terminal ends of inteins which is always Cys, Ser or Thr (Figure 3z, position 9), is not conserved, since only a few Type A BIL domains have a serine or threonine residue in that position. Other Type A BIL domains have aspartate, glutamate, asparagine, tyrosine or alanine residues in that position, which are not found in any intein.

Type B BIL domains were also found to share 6 Hint-like consensus sequence motifs (Figures 3h-i and 3l-o) and two novel motifs (Figures 3j-k) having no known Hog protein or intein Hint domain counterpart. Type B BIL domain carboxy terminal ends were found to have a conserved position comprising Cys, Ser or Thr residues (Figure 3o, position 6), potentially corresponding to the carboxy terminal flanking position of inteins (Figure 3z, position 9), however this carboxy terminal residue of Type B BIL domains (Figure 3o, position 7) is not preceded by the His-Asn motif typically found in inteins (Figure 3z, positions 7–8) and in Type A BIL domains (Figure 3g, positions 6–7). The -SH/-OH groups on the side chains of the aforementioned Cys, Ser or Thr residues in intein host proteins have been found to be essential for ligation of the intein carboxy and amino flanks in the protein splicing reaction (Xu MQ. and Perler FB., 1996. EMBO J. 15, 5146).

In Type A BIL domains whose carboxy terminal ends are not flanked by Thr

or Ser residues, Asn cyclization may nevertheless occur without trans-esterification by the flanking residue. Alternately, trans-esterification may occur by the mildly nucleophilic residues found in this position. In the first case the BIL domain would be cleaved from a segment flanking its carboxy terminal end, and in the second case protein splicing would occur. Since Type B BIL domains do not have any conserved Asn or Gln residue at their carboxy terminal end, cleavage of this end could then proceed by a mechanism different from the Asn and Gln cyclizations of inteins (Paulus H., 2000. Annu Rev Biochem. 69, 447).

Key amino acid residues corresponding to protein splicing active sites (marked by asterisks in Figures 3u, 3w, and 3z–position 9) were found to be conserved in Type B BIL domains.

Both Type A and Type B BIL domains were found to be distinct from inteins in having additional unique sequence motifs, in not being integrated in highly conserved sites of essential proteins, and in not comprising endonuclease domains.

Phylogenetic distribution of BIL domains: The phylogenetic distribution of the BIL domains identified is shown in Table 2. BIL domains were identified in 3 evolutionarily distant bacterial types – alpha, beta and gamma proteobacteria (gram-negative bacteria), actinobacteria (high GC gram-positive bacteria), and *Bacillus/Clostridium* group bacteria (low GC gram-positive bacteria).

Both presence and genomic distribution were found to be variable, even in closely related species and strains. For example, 1, 3 and 6 ORFs encoding BIL domains were identified, respectively, in *N. meningitidis* strains whose genomes have been completely or almost completely sequenced; 2 and 14 ORFs encoding BIL domains were identified in 2 different *Rhodobacter* species; and while one such ORF was identified in *P. syringae*, none were found in *P. aeruginosa* and *P. putida*.

BIL domains and inteins were found to coexist in certain species. For example, the genome of *M. magnetotacticum* was found to comprise ORFs encoding both Type A and Type B BIL domains, and that of *T. fusca* was found to comprise ORFs encoding both BIL domains and inteins.

The variability observed in the number of BIL domain ORFs in different species is probably due to gene duplications. As shown in a dendrogram demonstrating phylogenetic relationships of Type A BIL domains (Figures 4a), all

BIL domains derived from *Neisseria* species cluster together, and BIL domains from different species sub-cluster as well, implying that all *Neisseria* BIL domains arose by duplication from a single ancestor and that some are paralogs within different species. The latter is corroborated by the apparent duplication of some gene loci containing BIL domains in these species (not shown). Clustering of BIL domains from the same species was also observed in *C. thermocellum* (Figure 4a) and *M. Magnetotacticum* (Figure 4b).

BIL domain host proteins: BIL domains were identified in putative ORFs coding for a few hundred to a few thousand amino acids. Several BIL domains were found to be flanked by domains present in secreted bacterial proteins. In *P. syringae* and *M. haemolytica*, BIL domains were identified near the carboxy terminal end of FhaB-like ORFs. FhaB is a very large *Bordetella* gene coding for a secreted filamentous hemagglutinin protein, which functions as an adhesin important for *B. pertussis* virulence (Smith AM. *et al.*, 2001. FEMS Microbiol Rev. 25, 309). Three of the *R. capsulatus* BIL domain-containing ORFs include RTX repeats – calcium binding repeats found in various secreted bacterial proteins, including many toxins (Coote JG., 1992. FEMS Microbiol Rev. 8, 137). In *N. meningitidis* and *N. gonorrhoeae*, BIL domains were identified in MafB proteins. These are part of multiple adhesin family possibly involved in glycolipid adhesion to cells (Naumann *et al.*, 1999. Curr Opin Microbiol 2, 62-70; Paruchuri *et al.*, 1990. Proc Natl Acad Sci U S A. 87, 333-7). Three other *Neisseria* BIL domains were found to have an HNH nuclease domain in amino acid sequences flanking their carboxy terminal ends. HNH domains are found in various DNase and endonuclease proteins including secreted toxins (Belfort M. and Roberts RJ., 1997. Nucleic Acids Res. 25, 3379; James R. *et al.*, 1996. Microbiology 142, 1569). A domain present in the amino acid sequence flanking the carboxy terminal end of a BIL domain in the gram-positive bacterium *T. fusca* is also found in a short, conserved *Salmonella* ORF (GenBank accession NP_454902) and in an amino acid sequence flanking the carboxy terminal end of a *N. meningitidis* FhaB/hemolysin protein (gene NMA0688). Both of these proteins are from gram-negative bacteria and are likely to be secreted.

Table 2. Phylogenetic distribution of BIL domains identified.

Taxonomic group	Species	Species code	BIL type	No. of BIL domains identified in organism
alpha proteobacteria	*Rhodobacter capsulatus* SB1003	rhoca	B	14*
alpha proteobacteria	*Rhodobacter sphaeroides* 2.4.1	rhosp	B	2*
alpha proteobacteria	*Silicibacter pomeroyi* DSS-3	silpo	B	16*
alpha proteobacteria	*Brucella melitensis* 16M/*Brucella suis*	brume	B	1
alpha proteobacteria	*Magnetospirillum magnetotacticum* MS-1	magma	A	1
alpha proteobacteria			B	5*
alpha proteobacteria	*Methylobacterium extorquens* AM1	metex	B	1*
alpha proteobacteria	*Rhizobium leguminosarum bv. viciae 3841*	rhile	B	1
beta proteobacteria	*Neisseria meningitidis* Z2491	neimeA	A	1
beta proteobacteria	*Neisseria meningitidis* MC58	neimeB	A	3
beta proteobacteria	*Neisseria meningitidis* FAM18	neimeC	A	6*
beta proteobacteria	*Neisseria gonorrhoeae* FA1090	neigo	A	6
beta proteobacteria	*Chromobacterium violaceum ATCC 12472*	chrvi	A	1
gamma proteobacteria	*Pseudomonas syringae* DC3000	psesy	A	1*
gamma proteobacteria	*Pseudomonas fluorescens* PfO-1	psefl-PfO-1	A	1*
gamma proteobacteria	*Pseudomonas fluorescens* PfSBW25	psefl-SBW25	A	1*
gamma proteobacteria	*Mannheimia haemolytica* PHL213	manha	A	1*
delta proteobacteria	*Myxococcus xanthus DK1622*	myxxa	A	3*
spirochaetes	*Leptospira interrogans 56601*	lepin	A	3
actinobacteria	*Streptomyces coelicolor A3(2)*	strco	A	1
actinobacteria	*Streptomyces avermitilis MA-468*	strav	A	3
actinobacteria	*Thermobifida fusca YX*	thefu	A	1*
Bacillus/Clostridium group	*Clostridium thermocellum ATCC 27405*	cloth	A	10*
planctomycetes	*Pirellula species 1*	pirsp	A	1
planctomycetes	*Gemmata obscuriglobus UQM 2246*	gemob	A	2*
cyanobacteria	*Gloeobacter violaceus PCC 7421*	glovi	A	7
verrucomicrobium	*Verrucomicrobium spinosum DSM 4136*	versp	A	3*
unknown	*Unknown*		B	2*

*Genome not fully sequenced – total number of BILs may be greater.

BIL domain-mediated auto-cleavage/auto-splicing activity:

Type A BIL domain-mediated auto-cleavage/auto-splicing activity in an in-vitro transcription/translation system: Electrophoretic analysis (Figure 5a) showed that protein products generated following *in-vitro* transcription/translation of the MBP-PsyBIL-CBD expression construct pC2C-PsyBIL displayed molecular weights

corresponding to the uncleaved precursor MBP-PsyBIL-CBD, the splicing product MBP-CBD, and the carboxy terminal cleavage product MBP-PsyBIL. Two additional protein products displayed molecular weights of 43 and 45 kDa. Control reactions using the MBP-CBD expression construct pC2C as a transcription template produced two protein products, one corresponding in weight to MBP-CBD and the other to the MBP portion thereof. The latter may be observed in chimeric proteins having MBP as an amino terminal tag (refer to: NEB instruction manual "pMAL protein fusion and purification system", Catalog #E8000S). The 43 kDa protein product may represent a premature transcription or translation stop side product unrelated to BIL domain-mediated activity. Appearance of the 45 kDa band, not seen in the control reaction and slightly larger than the expected weight of MBP, may be due to an additional termination point introduced in the BIL domain. As radioactive methionine was used to label the reaction products, and as, unlike the MBP and BIL domains, the CBD domain lacks a methionine residue, its isolated product cannot be visualized according to the protocol employed.

The relative amounts of the MBP-PsyBIL-CBD, MBP-PsyBIL and MBP-CBD protein products were found to be 15 %, 57 % and 28 %, respectively.

These results therefore demonstrated the capacity of Type A BIL domains to auto-cleave and auto-splice flanking sequences.

Type B BIL domain-mediated auto-cleavage and auto-splicing activity in an in-vitro transcription/translation system: Electrophoretic analysis (Figure 5b) showed that protein products generated following *in-vitro* translation of the MBP-RspBIL2-CBD expression construct pC2C-RspBIL2 included proteins with sizes corresponding to the unprocessed MBP-RspBIL2-CBD precursor, the carboxy terminal cleavage product MBP-RspBIL2, the MBP-containing amino terminal cleavage product, and the splicing product MBP-CBD. Also apparent was a 43 kDa MBP-containing fragment which also appeared in the control reaction, as described above.

These results therefore demonstrated the capacity of Type B BIL domains to auto-cleave and auto-splice flanking sequences. The putative protein splicing motifs and catalytic residues of the FhaB_psesy amino acid sequence responsible for the observed autoprocessing activity are shown in Figure 5c (SEQ ID NO: 107).

Recombinant BIL domains expressed in-vivo in E. coli display auto-cleavage and auto-splicing activity – mass spectrometric confirmation of BIL-mediated autoprocessing activity: In order to examine the possibility of industrially producing functional recombinant Type A BIL domain, pC2C-PsyBIL was overexpressed *in-vivo* in *E. coli*, and the recombinant protein products were analyzed for BIL domain-mediated autoprocessing activity.

Similarly to the *in-vitro* results described above, purified protein from *E. coli* transformed with pC2C-PsyBIL was found to include a protein product having a molecular weight corresponding to the MBP-PsyBIL carboxy terminal cleavage product, as well as a protein product having a molecular weight corresponding to the MBP-CBD auto-splicing product, as determined by both Coomassie Blue staining and Western immunoblotting analysis of SDS-PAGE separated proteins (Figures 6a and 6b, respectively). The main product was again observed to be MBP-PsyBIL protein, as displayed by comparing the quantity thereof produced to that of the MBP-CBD protein product when both were purified using amylose beads (Figure 6a, lane 3).

The identities of the PsyBIL reaction products, the MBP-CBD and MBP-BIL protein products were also confirmed by mass spectrometry analysis (Figures 6c and 6d, respectively). The measured mass of the MBP-CBD protein (50,602.07 Da) was found to be in close agreement to the expected mass of the unmodified protein (50,266.39 Da). The measured and expected masses of MBP-BIL protein were also found to be in close agreement: 59,332.79 and 59,070.11 Da, respectively. A prominent peak with a mass of 43,303 Da, was also observed in MALDI spectra of electroeluted 50 kDa MBP-CBD protein. Such cross-contamination can be observed in gel purified protein bands (A. Shainskaya, unpublished). Reactivity of the 43 kDa band with anti MBP tag antibody (Figure 6b) indicates that this band is a truncated product.

Peptide mass mapping of the 50.1 and 59.3 kDa protein products proteins by MALDI analysis (Tables 3 and 4, respectively) confirmed their assigned identity as MBP-CBD (Figure 7; SEQ ID NO: 108) and MBP-BIL (Figure 8; SEQ ID NO: 109), respectively, in particular by identifying the splice junction of the MBP-CBD protein (peptide position 388–396; Figure 7; Table 3) and the carboxy terminal end of the MBP-BIL protein (peptide position 535–541; Figure 8; Table 4) with accuracies of 27

and 100 ppm, respectively.

Thus, following extensive experimentation, the splicing junction and cleavage points were found to precisely correspond to those predicted from the sequence similarity of the BIL and intein domains, thereby unambiguously demonstrating BIL domain autoprocessing capacity.

Table 3. MALDI identification of peptides of the 50.1 kDa MBP-CBD splicing product of MBP-PsyBIL-CBD.

Peptide position*	$[M+H]^+$ calculated mass (Da)	$[M+H]^+$ measured mass (Da)	Mass accuracy (ppm)
1–2	278.1538	278.1460	28
3–7	563.2677	563.2599	13
8–16	1057.604	1057.596	7
8–26	2047.363	2047.35	6
27–35	1064.532	1064.586	50
28–35	936.442	936.491	52
28–30	423.2244	423.215	22
90–99	1267.647	1267.6	37
129–138	1201.522	1201.602	66
129–141	1571.732	1571.823	57
191–201	1188.642	1188.711	58
172–180	1129.55	1129.57	17
253–274	2137.972	2138.147	81
279–296	2095.812	2096.030	104
297–306	1010.472	1010.612	138
328–345	2109.02	2109.017	1
356–387	3461.516	3461.33	53
364–387	2576.53	2576.41	46
388–396	983.48	983.55	71
397–435	3986.34*	3986.13	52
397–438	4380.80*	4379.90	205
439–461	2634.93*	2634.72	79

* mass corresponds to peptide with an alkylated cysteine residue

These results therefore fully and clearly demonstrate the capacity of BIL domains to auto-cleave and auto-splice flanking sequences, similarly to inteins. The presently described BIL domains therefore represent a novel and highly useful class of autoprocessing proteins which can be harnessed for manipulating and modifying proteins, for example as depicted in Figure 9. These results furthermore demonstrate

the suitability of utilizing genetically transformed host cells, such as *E. coli*, to industrially express chimeric proteins which comprise functional BIL domains.

Table 4. MALDI-identification of carboxy terminal peptides of the 59.3 kDa MBP-PsyBIL carboxy terminal cleavage product of MBP-PsyBIL-CBD.

Peptide position	$[M+H]^+$ calculated mass (Da)	$[M+H]^+$ measured mass (Da)	Mass accuracy (ppm)
535–541	897.4947	897.47	27
537–541	656.3156	656.25	100

Discussion: BIL domains are present in several hyper-variable bacterial proteins, such as FhaB adhesins and MafB proteins of *Neisseria* strains. Their immediate flanks are the most variable portions of the proteins and they themselves are not always present in these proteins, even in closely related strains of the same species. Some, and perhaps all, proteins with BIL domains seem to be secreted proteins. BIL domains might enhance the variability of secreted proteins by their protein splicing and cleavage activity as detailed below.

As described above, the amino terminal ends of BIL domains, and of Hint domains of inteins and Hog proteins are very similar (Figures 3a-z). Thus all these domains probably form labile ester bonds on their amino terminal ends. In proteins with BIL domains these ester bonds could be attacked by various nucleophilic molecules, including peptides, proteins and small reactive compounds, such as glutathione or cysteine. Such reactions would ligate the attacking nucleophiles to a carboxy terminal position of the host protein and release the BIL domain and the host protein region downstream to it. This is analogous to Hedgehog protein maturation where the Hint domain mediates the attachment of a cholesterol molecule to the cleaved Hedge domain. In adhesins with BIL domains this putative ligation might serve to covalently attach the bacteria to its adhesion target. Additionally, released BIL and carboxy terminal domains could have a function of their own. For example, in pathogenic bacteria that have such proteins, the released domains could serve as decoys to the immune system.

In *Neisseria* strains, sequences encoding BIL domains appear as either as short open reading frames downstream of MafB genes and in the carboxy terminal ends of

these proteins upstream of a variable domain. This suggests that, at least in such *Neisseria* strains, BIL domains function as cassettes which can be fused to genes by genetic rearrangement to promote the variability of the encoded proteins. Other microevolutionary processes in *Neisseria* and *Ralstonia solanacearum*, a plant pathogen bacterium with a wide host range, are known to generate different carboxy terminal ends for surface-exposed and virulence proteins (Parkhill *et al.*, 2000. Nature, 404, 502-506; Salanoubat *et al.*, 2002. Nature 415, 497-502).

Not all species with BIL domains are pathogens and many pathogenic bacteria with fully sequenced genomes do not have BIL domains. BIL domains might be used in different processes not connected with pathogenicity. For example, BIL domain activity might be one way for bacteria to attach to diverse surfaces.

In summary, two novel types of Hint domain-containing proteins, BIL Types A and B, were identified. Both types have the active site sequence features of the Hint domains but also possess sequence features that distinguish them from the known Hint domains and from each other. BIL domains appear in different proteins from diverse bacteria, including pathogenic species of humans and plants, such as *Neisseria meningitidis* and *P. syringae*. These domains are present in variable protein regions and are typically flanked by domains that also appear in secreted proteins such as filamentous hemagglutinin and calcium binding RTX repeats. Phylogenetic and genomic analysis of BIL domain sequences suggests that they were positively selected for in different lineages. Type A and Type B BIL domains were cloned and shown to display auto-cleavage and auto-splicing of flanking polypeptide sequences in an *in-vitro* transcription/translation system, as well as when overexpressed in *E. coli*, thereby indicating the capacity of BIL domains to autocatalyze post-translational modifications of host proteins.

Conclusion: The above-described experimental results demonstrate the capacity of the autoprocessing polypeptides of the present invention to efficiently auto-cleave and auto-splice flanking sequences. The presently described experimental results furthermore demonstrate the feasibility of utilizing genetically transformed host cells, such as *E. coli*, for efficient industrial production of such autoprocessing polypeptides. Thus, the autoprocessing polypeptides and the polynucleotides encoding such polypeptides of the present invention significantly expand and enhance

the available repertoire of available autoprocessing polypeptides having utility in numerous commercially important protein engineering applications, such as protein purification, affinity selection of display phages and post-translational protein ligation.

EXAMPLE 2

Supplementary evidence demonstrating C-terminal in-vitro and in-vivo autocleavage by chimeric protein including the Type B BIL domain BIL2_rhosp

Materials and Methods:

In order to analyze the capacity of Type B BIL domain BIL2_rhosp (Table 1, Figure 1b) to display autoprocessing activity, genetic sequences encoding this BIL domain including one flanking amino acid residue at each terminus were cloned for expression as a chimeric protein tagged at its amino terminal end with the *malE* gene-encoded maltose-binding protein (MBP) affinity tag, and at the carboxy terminal end with the *B. circulans cbd* gene-encoded chitin-binding domain (CBD) affinity tag. The resultant "MBP-RspBIL2a-CBD" chimera was expressed *in-vivo* and *in-vitro* and resulting protein products were analyzed for evidence of BIL domain-mediated autoprocessing activity according to methods described in Example 1 above.

Experimental Results:

The *in-vivo* expressed MBP-RspBIL2a-CBD chimera was shown to display C-terminal auto-cleavage activity via amylose-based affinity purification, electrophoretic separation, and Coomassie blue staining of the electrophoretically separated proteins (Figure 10a). The *in-vivo* expressed MBP-RspBIL2a-CBD chimera was also shown to display C-terminal cleavage via Western immunoblotting analysis of SDS-PAGE separated proteins (Figure 10b). The *in-vitro* expressed MBP-RspBIL2a-CBD chimera was shown to display C-terminal cleavage following [^{35}S]-methionine–labeling and autoradiography of electrophoretically separated protein (Figure 10c). Evidence from intact mass mass-spectrometry of the MBP-BIL specific carboxy terminal cleavage product indicates that the C-terminal end of this product is located 6-11 amino acid residues from the predicted carboxy terminal end of the BIL domain towards the N-terminus. Evidence for the MBP-RspBIL2a identity of the cleavage product was obtained from Western Blot, and affinity column analysis. The presence

of the protein chaperone DnaK was detected during affinity purification of the BIL products. This chaperone may bind the BIL domain, and may also be involved in its activity.

Conclusion: There is ample evidence that a chimeric protein which comprises the type B BIL domain BIL2_rhosp has the capacity to efficiently display carboxy terminal auto-cleavage activity. Hence, such a BIL domain can therefore be advantageously exploited in applications benefiting from an auto-cleaving chimeric protein.

EXAMPLE 3

N-terminal autocleavage of in-vivo expressed chimeric polypeptide comprising the Type B BIL domain 4825_rhosp

Materials and Methods:

In order to analyze the capacity of Type B BIL domain 4825_rhosp (Table 1, Figure 1b) to display autoprocessing activity, genetic sequences encoding this BIL domain with 14 amino terminal-flanking and 51 carboxy terminal-flanking amino acids were cloned for expression as a chimeric protein tagged at its amino terminal end with the *malE* gene-encoded maltose-binding protein (MBP) affinity tag, and at the carboxy terminal end with the *B. circulans cbd* gene-encoded chitin-binding domain (CBD) affinity tag. The resultant "MBP-4825rhosp-CBD" chimera was expressed *in-vivo* and resulting protein products were analyzed for evidence of BIL domain-mediated autoprocessing activity according to methods described in Example 1 above.

Experimental Results:

The *in-vivo* expressed MBP-4825rhosp-CBD chimera was shown to display N-terminal auto-cleavage activity via chitin or amylose based affinity purification of expressed protein and electrophoretic separation, and Coomassie blue staining of electrophoretically separated protein (Figure 11a). The chimera was also shown to display N-terminal cleavage via Western immunoblotting analysis of SDS-PAGE separated proteins using anti-CBD and anti-MBP antibodies (Figure 11b). The cleavage site was found to be located exactly between the predicted N-terminal residue of the BIL domain and the amino terminal flanking sequence as demonstrated

by amino terminal end sequencing. Large amounts of the chaperone GroEL were detected during the purification of the protein products when overexpressing the MBP-4825rhosp-CBD chimera the whereas no GroEL was detected when overexpressing control chimera lacking the BIL domain.

Conclusion: A chimeric protein which comprises the type B BIL domain 4825_rhosp has the capacity to efficiently display N-terminal auto-cleavage activity. Hence, such a BIL domain can therefore be advantageously exploited in applications benefiting from an auto-cleaving chimeric protein.

EXAMPLE 4

Auto-splicing and C-terminal auto-cleavage of in-vivo expressed chimeric protein including the Type A BIL domain BIL4_cloth

Materials and Methods:

In order to analyze the capacity of Type A BIL domain BIL4_cloth (Table 1, Figure 1a) to display autoprocessing activity, genetic sequences encoding this BIL domain were cloned, including one flanking amino acid residue of the host protein at the C-terminal end, for expression as a chimeric protein tagged at its amino terminal end with the *malE* gene-encoded maltose-binding protein (MBP) affinity tag, and at the carboxy terminal end with the *B. circulans cbd* gene-encoded chitin-binding domain (CBD) affinity tag. The resultant "MBP-BIL4cloth-CBD" chimera was expressed *in-vivo* and resulting protein products were analyzed for evidence of BIL domain-mediated autoprocessing activity according to methods described in Example 1 above.

Experimental Results:

The *in-vivo* expressed MBP-BIL4cloth-CBD chimera was shown to display auto-splicing activity (MBP-CBD specific fragment) as shown by Western immunoblotting assay using anti CBD and anti MBP antibody probes (Figures 12a and 12b, respectively), and as shown by Coomassie blue staining of electrophoretically separated protein products isolated via amylose based or chitin based affinity chromatography depicting auto-splicing activity. The identity of the auto-splicing product was verified via mass-spectrometry peptide mapping (data not shown). Carboxy terminal auto-cleavage activity (MBP-BIL domain specific

fragment) was also detected by Western immunoblotting assay using an anti MBP antibody probe (Figure 12b) and by Coomassie blue staining of electrophoretically separated protein products isolated via amylose based affinity chromatography (Figure 12c).

Conclusion: A chimeric protein which comprises A type a BIL domain, such as the type A BIL domain BIL4_cloth, has the capacity to efficiently display auto-splicing and carboxy terminal auto-cleaving activity. Hence, such a BIL domain can therefore be advantageously exploited in applications benefiting from an auto-splicing and/or carboxy terminal auto-cleaving chimeric protein.

It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable subcombination.

Although the invention has been described in conjunction with specific embodiments thereof, it is evident that many alternatives, modifications and variations will be apparent to those skilled in the art. Accordingly, it is intended to embrace all such alternatives, modifications and variations that fall within the spirit and broad scope of the appended claims. All publications, patents, patent applications and sequences identified by their accession numbers mentioned in this specification are herein incorporated in their entirety by reference into the specification, to the same extent as if each individual publication, patent, patent application or sequence identified by their accession number was specifically and individually indicated to be incorporated herein by reference. In addition, citation or identification of any reference in this application shall not be construed as an admission that such reference is available as prior art to the present invention.